SciDAC Visualization and Analytics Center for Enabling Technologies
Semi-Annual Progress Report
December 2009 through May 2010

E. Wes Bethel[*] and Chris Johnson[†]
Principal Investigators

Charles Hansen, Valerio Pascucci, Claudio Silva, Allen Sanderson, Tom Fogal, Kristi Potter, Attila Gyulassy, Brian Summa, Josh Lavine, Sidarth Kumar, Giorgo Scorzelli, Abhishek Tripathi, Julien Tierney, Lauro Lins, Juliana Freire, Emanuele Santos, Harsh Bhatia, Guoning Chen, Shreeraj Jadhav, Luis Nonato [‡]

Sean Ahern, George Ostrouchov, Dave Pugmire, Jeremy Meredith[§]

Eric Brugger, Peer-Timo Bremer, Brad Whitlock[¶]

Ken Joy, Eduard Deines, Christoph Garth, David Camp, Bernd Hamann[‖]

Hank Childs, Janet Jacobsen, Prabhat, Oliver Rübel, Daniela Ushizima, Gunther Weber[**]

May 2010

[*]Lawrence Berkeley National Laboratory
[†]Scientific Computing Institute, University of Utah
[‡]Scientific Computing Institute, University of Utah
[§]Oak Ridge National Laboratory
[¶]Lawrence Livermore National Laboratory
[‖]University of California – Davis
[**]Lawrence Berkeley National Laboratory

Contents

1 Executive Summary

1.1 Overview

The SciDAC Visualization and Analytics Center for Enabling Technologies (VACET) focuses on leveraging scientific visualization and analytics software technology as an enabling technology for increasing scientific productivity and insight. Our mission is to foster scientific insight through creating and deploying effective data understanding technology that is truly responsive to the needs of our stakeholders in the scientific research community who are "awash in data." It is widely accepted that one of the bottlenecks in contemporary science is the need to gain insight from vast collections of complex data.

The vision for our Center is to respond directly to this challenge by adapting, extending, creating when necessary and deploying visualization and data understanding technologies for our science stakeholders. Organized as a Center for Enabling Technologies, we are well positioned to be responsive to the needs of a diverse set of scientific stakeholders in a coordinated fashion using a range of visualization, mathematics, statistics, computer and computational science and data management technologies.

We are pleased to report accomplishments during the period of December 2009 through May 2010, both in terms of impact for scientific stakeholders and in terms of providing leadership in the visualization and analysis community.

1.2 Accomplishments

Science Application Projects

Accelerator Modeling. Our team has developed and deployed new methods for accelerating visual data analysis and exploration through a combination of parallelism and leveraging search/index technology from the SciDAC Scientific Data Management Center aimed at increasing understanding of how high energy beams form in laser-wakefield accelerators (Section 2.2). We extend this work to support statistical analysis of time-evolving beamlines in single simulation runs with the aim of enabling comparative analysis across across multiple simulation runs (ensembles) (Section 2.3). It is worthy of note that our work in designing and implementing "Named Selections" in VisIt, along with the integration of VisIt and FastBit, is an accomplishment we are leveraging to make progress in Fusion projects. This work relies on technology from one of our fundamental infrastructure projects, integration of FastBit index/query into VisIt (Section 4.3).

Nuclear Energy. The Nek5000 code team at ANL rely on VisIt for their visualization needs and on VACET for expert assistance for solving difficult visual data exploration and analysis problems. We recently performed an in-depth analysis, which is part of a code validation effort, that computes the "residence time" of air inside a box. Our approach was to use our integral curve computation algorithm to advect particles through space then to compute statistics (residence time) on those particles (Section 2.7).

Computational Chemistry. The MADNESS code, which is supported by multiple DOE and SciDAC efforts, produces a form of output that is not readily accessible to contemporary visualization tools. Our team is undertaking an effort to create the data loaders necessary to import this data into VisIt, thereby making visual data exploration and analysis capabilities accessible to a number of DOE computational chemistry projects that rely on the MADNESS code (Section 2.8).

Combustion. Our team has pioneered the use of topological analysis techniques aimed at helping to shed light on the relationship between turbulence and combustion characteristics in laboratory-scale simulations of lean, pre-mixed combustion 2.5. Working with DNS simulation

output, we are investigating techniques based upon topological analysis for studying the shape characteristics of structures associated with combustion exinction and re-ignition 2.6.

Fusion. We are focusing on several interrelated efforts aimed at helping solve data understanding problems in the Fusion science community. One effort is producing new algorithms for robustly computing high-resolution Poincaré plots, a commonly used technique for studying and analyzing magnetic island formation in plasmas (Section 2.9). Another is to leverage query-driven visualization technology, successfully deployed to the accelerator science community, to meet fusion-specific data and science understanding needs (Section 2.10).

Technology Incubation Projects

Hybrid parallelism – Volume Rendering. Our team conducted a study to better understand performance and scalability limits of a staple visualization algorithm, raycasting volume rendering, at extreme concurrency on a large, multi-core platform using hybrid parallelism, a blend of distributed- and shared-memory parallelism (Section 3.1). The results indicate the hybrid parallelism approach runs faster, uses less memory and communication bandwidth than traditional approaches. These results are significant because they pave the way for future visual data exploration and analysis work to be able to take full advantage of petascale and beyond platforms. This field-leading work won the Best Paper Award at the Eurographics Parallel Graphics and Visualization Symposium held in Norrköping, Sweden, May 2010.

Hybrid parallelism – Streamlines. In the same vein, our team explored performance characteristics of traditional and hybrid-parallel implementations of another staple visualization algorithm, streamlines (Section 3.2). The hybrid parallel approach also proved to have distinct performance advantages in this case over the traditional, MPI-based approach.

Algorithm: Calculating Integral Curves. Here, our aim is to enable physicists, chemists and fluid dynamicists to visualize and analyze their state-of-the-art simulation flow-field data using robust, efficient and scalable integration-based visualization techniques over a wide variety of data representations and problem characteristics. Our solution is based on a novel code framework, integrated into VisIt for widespread dissemination to the science community, that enables the efficient and scalable computation of integral curves in vector fields represented over regular, structured, unstructured and AMR meshes (Section 3.4). Based on our framework, scientists will be able to leverage modern integration-based visualization techniques to visualize and analyze vector field data to investigate phenomena such as transport and mixing. Our work specifically addresses very large / petascale data, to enable robust analysis on current and future-generation datasets (Section 3.3).

Software Engineering and Infrastructure

Production-quality, Adaptive Mesh Refinement Visualization. Our team continues to work closely with the SciDAC Applied Partial Differential Equations Center (APDEC) to maintain and apply VACET's production-quality AMR visualization technology to the SciDAC community. Ongoing efforts include software engineering to add new features, fix bugs, and improve performance. Recent accomplishments include a cover image on SciDAC Review (Section 4.2).

Data-parallel Statistical Analysis. An ongoing effort in VACET is to bring to bear the power of the R statistical analysis package, which is effectively limited to operation on a single core, on large-data problems and on DOE's large parallel computational platforms. To achieve that objective our team is performing software engineering to the Rmpi package to enable use of R in a data-parallel mode (Section 4.4).

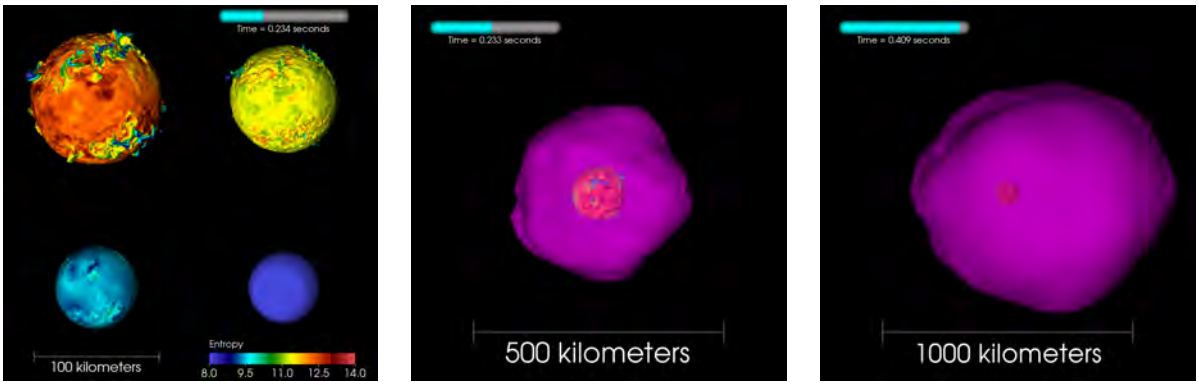
Production-quality, Petascale Visualization a Reality. The VACET team has conducted a performance study of the VisIt visualization application to better understand its performance at very high levels of concurrency and on very large data sets. These experiments consist of a weak scaling study of a production-quality visualization code (VisIt) on six different platforms on the largest-ever data sets published in open literature. This result is significant because it shows that VACET's production-quality visualization software is capable today of handling tomorrow's scientific simulation data sets on DOE's largest computational platforms (Section 4.1).

2 Specific Stakeholder Projects

2.1 Astrophysics: The Community Astrophysics Consortium

The Problem. To ensure that the visualization and analysis needs of the astrophysics community, and specifically the Computational Astrophysics Consortium (CAC), are met. This involves (i) providing them tools for bread-and-butter functionality, (ii) providing support, (iii) helping with high end movies, and (iv) helping with high-end analysis.

The Solution. VACET is providing and deploying to the CAC with production quality visualization tools to meet day-to-day needs. We are also providing support to the CAC by fielding questions that come up in solving specific science problems, as well as providing in-depth consulting for creating "difficult" images and movies as well as helping to devise new methods for high-end analysis.



(a) Four different isodensities showing activity at various depths in the supernova core.

(b) Isosurface of outer shock, with core of supernova colored (earlier timestep).

(c) Isosurface of outer shock, with core of supernova colored inside (later time). The asymmetry of the surface was of high interest to the scientists. Also, the code to identify the shock required custom VisIt development.

Figure 1: Example images of science results produced by CAC members.

The Impact. We are successfully delivering a production-quality visual data exploration and analysis tool to the SciDAC astrophysics community, thereby enabling new ways of exploring and understanding scientific data as well as enabling cost savings through reduced duplication of effort (they don't have to build or buy a tool to meet their visualization needs). Visualization results have been used by CAC members in presentations and publications that describe new science.

2.2 Accelerator Modeling – Beam Analysis

The Problem. Traditionally, physicists perform a manual classification of "particle bunches" through visual inspection of large amounts of simulation output data. The objective is to find those particles that are undergoing acceleration in a laser wakefield. Such manual inspection is time consuming. One main goal of this project is reduce the time for detection and analysis of particle beams.

Currently, particle bunches are classified based on a single reference timestep. The project aims to use the complete temporal history of the particles to enable more accurate beam classification. The large sizes (TBs) of current datasets presents challenges for the analysis of the complete time series. Development of efficient methods for data analysis and data management methods (such as FastBit) is, therefore, a central part of this project. Simulations have largely varying temporal and spatial resolution which analysis methods need to reliably handle.

The Solution. First, we address the manual search problem by developing new algorithms that perform automatic particle beam detection by locating particles undergoing acceleration in large, time-varying accelerator simulation results. We solve the temporal analysis problem with new algorithms that perform automatic detection across time. Our solutions rely on new algorithms for computing 3D histograms in FastBit[1] These methods provide us with information about the histogram counts as well with bitvectors allowing us to quickly access particles associated with a set of histogram bins. This functionality is essential to enable efficient implementation of the beam path analysis algorithm.

The Impact. Automating the detection of particle beams significantly reduces the time required for manual inspection of the data. This step is essential to enable analysis large collections of simulation data. By considering information from the complete time series we enable more accurate classification and analysis of particle beams. Analysis of the temporal evolution of particle bunches is essential for the understanding of how particles are accelerated in a laser wakefield accelerator. In the long run, this capability will certainly help accelerator scientist to accelerate data understanding. We are also planning to integrate this method with other Visualization/Analysis tools (VisIt) as part of a broad set of high-performance tools for scientific knowledge discovery.

Accomplishments This Period.

Based on the knowledge we gained from our previous work, in this reporting period we focused on generalizing the algorithms to the problem of detecting all possible relevant acceleration structures. This type of approach is a compromise between a fully automatic and purely manual analysis. We automatically extract information about the main relevant acceleration structures and provide tools that allow scientists to quickly explore the found structures and decide which structures are most important.

This result is important in particular to be able to handle simulations with largely varying simulation settings. In practice different setups may be used to achieve high acceleration (e.g., single or multiple laser pulses). Depending on the simulation setup, the simulation may show different acceleration behavior and we need to be able to deal with these robustly.

In order to understand the general acceleration behavior of a simulation we need to be able to look at all possible relevant particle bunches not just the main beam(s). In cases were the behavior of a particular physical setup is not well-understood yet, this gives the physicists more flexibility and allows them to decide which features are most important rather than having the algorithm make (a possibly wrong) decision for them. A fully automatic analysis approach (as we described last reporting period) is important to support analysis of large collections of LWFA simulations.

[1] FastBit is index/query software from the SciDAC Scientific Data Management Center.

Figure 2: Massively parallel VORPAL simulations of LOASIS (LBNL) experiments show the structure of a plasma density wave, or wake (blue), driven by an intense laser pulse traveling upward through the image. Simulation by C.G.R. Geddes (LBNL). VisIt visualization by G.H. Weber and C.G.R. Geddes (LBNL).

A more exploratory approach is important in particular to analyze novel simulations that are not well understood.

2.3 Accelerator Modeling – Bunch Exploration

The Problem. Physicists exploring laser-wakefield accelerator modeling may use different setups high acceleration, e.g., single or multiple laser pulses). Depending on the simulation setup, the simulation may show different types of characteristics of acceleration. Therefore, there is the need to understand acceleration behavior resulting from different configurations of input parameters. Further, physicists desire to be able to compare results from multiple simulation runs to find optimal simulation settings, and to understand how different settings effect the acceleration behavior. Analysis of a large number of simulations is time consuming, making a purely manual analysis of the data prohibitively expensive. They need, therefore, analysis tools that support efficient analysis of simulation results with largely varying simulation settings. The analysis should minimize the time it takes to investigate simulation results, deal robustly with a large range of physical settings, deal robustly with 2D and 3D simulations, and simulations with varying temporal and spatial resolution, and be computationally efficient.

The Solution. We are exploring a semi-automatic analysis approach. We first analyze the data to detect all main acceleration structures (bunches). For each bunch we derive additional information to, e.g,. describe its quality via statistics and characterize its temporal evolution. We then provide the user with an overview of all the structures and allow the user to query all bunches

to find those that are most relevant for the current analysis. This approach, largely reduces the time for manual interaction while at the same time allowing scientist to define which bunches are of most interest. The analysis provides researchers with a summary of the acceleration behavior of a simulation, providing detailed insight of the general acceleration behavior.

The Impact. The analysis will reduce the time needed to detect and analyze a particle dataset. Our approach provides physicists more detailed insight of the general acceleration behavior than has been possible before. Our approach will enable physicists to analyze and compare results from simulations having widely varying initial conditions and parameters. This capability is essential in order to help physicists understand and improve particle accelerator designs.

Accomplishments this Period.

Based on the knowledge we gained from our previous work on automatic detection of particle beams (Section 2.2) we developed a general algorithm for detecting "all" possibly relevant acceleration structures (particle bunches). We automatically extract information about the main relevant acceleration structures and provide tools that enable physicists to quickly explore these structures and decide which are the most important. This type of approach is a compromise between a fully automatic and purely manual analysis. Rather than trying to automatically decide which particle bunches are most important, we extract detailed information (e.g., statistics) about each bunch. The physicist can then query the bunches to find those that are most important for the current analysis This approach allows us to significantly accelerate the analysis process while allowing the physicists to control the analysis process. By providing an overview of all main bunches, we allow physicists scientist to get an understanding of the general acceleration behavior.

In this reporting period we developed:

- A novel algorithm for detecting and classifying particles bunches in LWFA simulations.
- A graphical user interface that allows researchers to quickly explore and query the found bunches as well as to define which particles belong to a selected bunch of interest.
- We linked the analysis to VisIt in two ways. First, we create VTK files of particle traces which include detailed information about the bunches and that can be visualized in VisIt. Second, we support the creation of "Named Selections" that can be applied to any visualization in VisIt. Named selections contain a list of ID's describing which particles are selected (here a set of particles forming a particular bunch).

The Algorithms. The algorithm works in a pipeline fashion At each step of the pipeline, additional, optional filtering operations are provided to reduce the amount of particles or bunches under consideration. Most filtering options are also available in the GUI. The following provides an overview of the algorithm.

- Perform a "cumulative query." Execute a base query (e.g., $px > 1e10$) at each timestep and count for each particle how often it satisfied the query condition. Optional Filtering: Remove all particles that satisfy the condition less than m times.
- Trace the detected particles over the complete time series. Optional Filtering: Remove short traces.
- Along each particle trace, compute the local maxima in acceleration (i.e, local maxima in px over time/acceleration direction x). Optional filtering: Remove local maxima with only short duration.
- At each timestep the particles that reach a local maximum in acceleration are usually densely grouped. We compute at each timestep a 3D histogram (x,y,px) and extract the connected components of it. Each such group of particles is said to define a characteristic acceleration structures (i.e, a particle bunch or a characteristic substructure of it).

- For each acceleration structure compute a reference trace.
- Based on the reference trace compute the temporal phases of the bunch (formation, acceleration, deceleration, etc.) and define for each particle (i.e., each particle that sufficed the cumulative query) its distance to all acceleration structures. As in the beam path analysis, we here use two distance functions, one in physical and one in momentum space. We then also derive additional statistics to provide information about the estimated quality (e.g., compactness in physical and momentum space) for each bunch.
- Based on the distance fields we can then optionally, also compute a clustering of the data, i.e., we assign each particle to one acceleration structure (or to the background if it is not close enough to any relevant structure).

The User Interface. The user interface serves several purposes:

- We provide an overview of all found acceleration structures and their properties.
- We allow the user to query the acceleration structures based on their properties to allow the user to quickly identify the most important bunches (e.g., only those with a particular peak energy).
- Once the user has identified the bunch(es) of interest we provide an interface that allows the scientist to quickly identify the particles that belong to the bunch. We can here directly point the user to the timestep where a bunch reaches its peak energy so that the user can quickly select the bunches based on properties at that timesteps as well as the derived distance fields over time.

2.4 Climate: Deploying Advanced 3D Visualization to the Climate Community through ESG/CDAT

The Problem. With the increasing reliance on large scale simulations to understand and predict the global climate, tools that facilitate the analysis and distribution of climate simulations are becoming evermore crucial. The climate community needs a small set of tools that is broadly applicable, standardized, collaborative, and reliable to fully exploit the existing simulation capabilities. The Climate Data Analysis Tools (CDAT) framework provides such a platform for many standard 2D visualization and analysis techniques but must be extended to include 3D visualization and more advanced feature based analysis techniques.

The Solution. The goal is support the efforts of the global climate community to better understand and predict climate. In particular, the focus is on developing advanced visualization and analysis capabilities and deploying them into the Climate Data Analysis Tools (CDAT) framework. Our mission includes the following elements: (1) Deploying advanced visualization capabilities into the CDAT tool and create a clear path for similar integration in other tools. (2) Extending the visualization software to incorporate domain specific requirements, data formats, and vector field visualization. (3) Supporting time-dependent and cross-dataset comparison, visualization and analysis. (4) Developing new analytic capabilities for climate data (first deployed into CDAT/VCDAT). (5) Developing a visualization and data analysis scenario for understanding of complex coupled phenomena such as the multi-scale dynamics the complete carbon cycle on earth.

The Impact. We are deploying new capabilities in CDAT and are thus extending one of the most used tool in the climate community. Even with the current prototypical tools we have created several high profile visualizations and animations. In particular, we have created an animation showing that while the lower atmosphere is warming the outer atmosphere is cooling. This is a strong indication that the warming effects are not caused by a change in solar activity but are

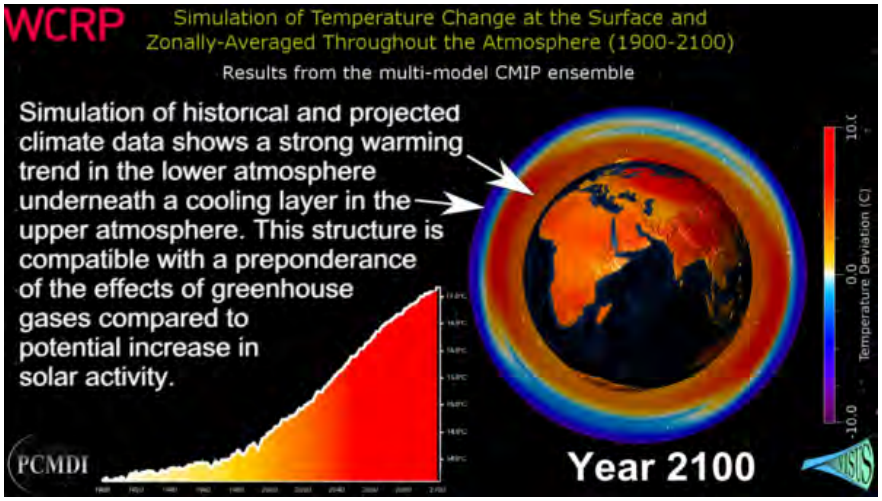


Figure 3: This image, one frame from the complete video which was shown in the WCRP booth at the climate meeting in Copenhagen, demonstrates the use of the visualization to communicate an important insight. The insight in this example is that the outer layer of the atmosphere is cooling while the lower layer is warming.

rather directly related to the insulating effects of carbon dioxide. This image, one frame from the complete video which was shown in the WCRP booth at the climate meeting in Copenhagen.

Future Plans. Work plans for the future include: (1) migrate the remaining FLTK interface to Qt; (2) move to XML delta-encoding for remote collaboration; (3) add vector field visualization to the existing code base; (4) prepare server to test and demonstrate remote data access.

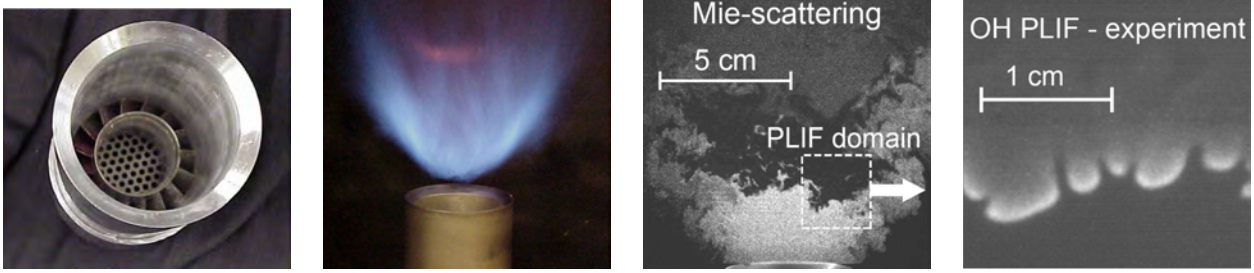
2.5 Combustion – Topological Analysis of Combustion Simulation Results on AMR Grids

The Problem.

Low-swirl injectors are emerging as an important new combustion technology. In particular, such devices can support a lean hydrogen-air flame that has the potential to dramatically reduce pollutant emissions in transportation systems and turbines designed for stationary power generation. However, hydrogen flames are highly susceptible to various fluid-dynamical and combustion instabilities, making them difficult to design and optimize. Due to these instabilities, the flame tends to arrange itself naturally in localized cells of intense burning that are separated by regions of complete flame extinction.

Existing approaches to analyze the dynamics of flames, including most standard experimental diagnostic techniques, assume that the flame is a connected interface that separates the cold fuel from hot combustion products. In cellular hydrogen-air flames, many of the basic definitions break down—there is no connected interface between the fuel and products, and in fact there is no concrete notion of a "progress variable" that can be used to normalize the progress of the combustion reactions through the flame. As a consequence, development of models for cellular flames requires a new paradigm of flame analysis.

Figure 4 shows the detail of a low-swirl nozzle. The annular vanes inside the nozzle throat generate a swirling component in the fuel stream. Above the nozzle the resulting flow-divergence provides a quasi-steady aerodynamic mechanism to anchor a turbulent flame. The middle figure illustrates such a flame for a lean premixed CH4-air mixture (the illustration shows a methane



(a) Photo of a typical laboratory low-swirl nozzle.

(b) Photo of a lean, premixed CH4 low-swirl flame.

(c) Experimental Mie scattering image of a lean premixed H2 flame

(d) PLIF data imaging the OH concentration in a lean premixed H2 flame.

Figure 4:

flame since H_2 flames do not emit light in the visible spectrum). The remaining images show typical experimental data from laboratory low-swirl, lean H2-air flames. Such data is used to extract the mean location and geometrical structure of instantaneous flame profiles. The images indicate highly wrinkled flame surfaces that respond in a complex way to turbulent structures and cellular patterns in the inlet flow-field.

Given the highly complex and dynamic nature of low-swirl flames analyzing them experimentally is challenging. Device-scale computer simulations of similar flames may provide more insight into the behavior of the flame, as well as ways to formulate and test new hypothesis. The Center for Computational Sciences and Engineering (CCSE) at Lawrence Berkeley Laboratory is using large, time-varying AMR combustion simulations to replicate low-swirl combustion and needs analysis tools to aid their understanding.

In particular, they are interested in the formation and evolution of burning cells, defined as contiguous regions of "high" fuel consumption. Furthermore, they are interested in the correlation between cell sizes and turbulence and various other statistical properties of the cells. An additional challenge is the fact that there exists no a-priori correct threshold of fuel consumption. In fact, determining a viable threshold and analyzing the stability of the results with respect to changes in the threshold is a primary goal of our analysis.

The Solution.

We are using topological techniques to analyze burning cells independent of a specific threshold. In particular, starting from the native AMR data we compute hierarchical merge trees of fuel consumption that encode a one-parameter family of segmentations with the fuel consumption threshold being the free parameter. Furthermore, we augment the segmentation with various pre-computed statistics (statistical moments etc.) of arbitrary species that, as the segmentation itself, can be accumulated on the fly into statistics for any specific fuel consumption threshold. By storing the merge trees hierarchically, we encode all possible cell segmentations in a single data structure. This data structure is two orders of magnitude smaller than the input, making it possible to explore interactively an entire family of features along with aggregate attributes, such as cell volume or average fuel consumption using pre-computed topological information.

Splitting segmentation information from the hierarchical merge trees, we create a lightweight index into the pre-segmented data that can be loaded on demand, thus enabling interactive analysis. A linked-view system uses this index to correlate the tracking graph with displays of segmentations supporting interactive exploration of their temporal evolution. Using pre-computed attributes, such as cell volume or average fuel consumption, it is possible to sub-select cells and explore the

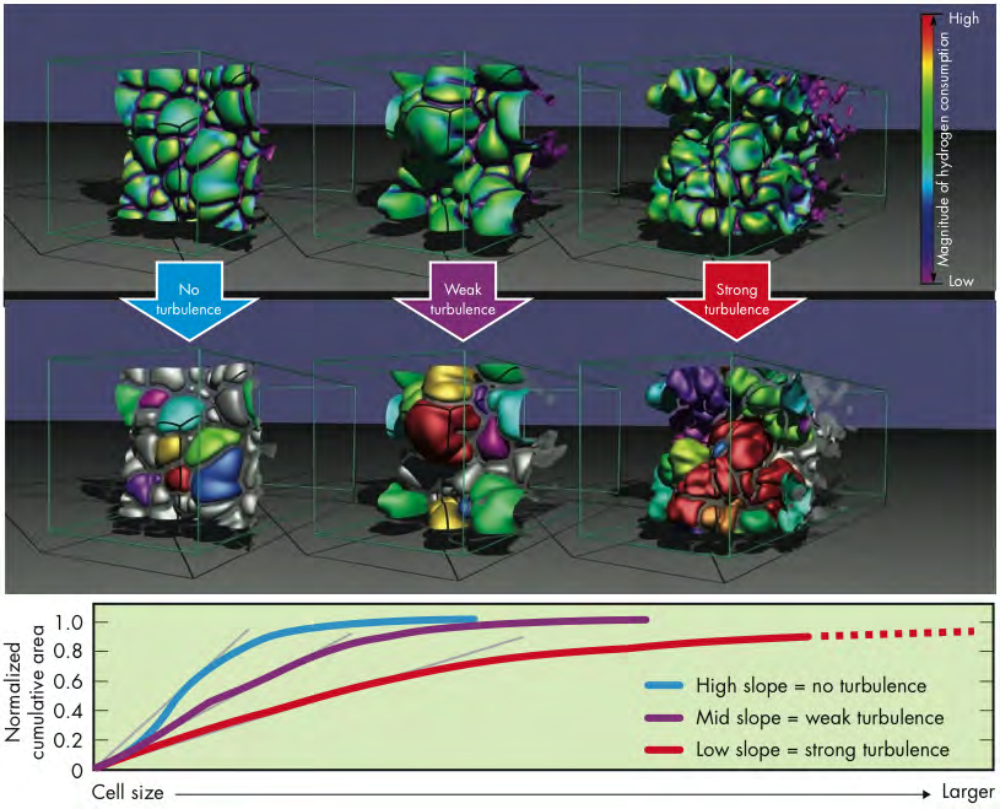


Figure 5: The top 3D diagrams show the flame intensity on the surface of cells at different turbulence levels, with the non-burning regions colored purple. For each turbulence level in the bottom 3D diagrams, a small set of burning cells are randomly colored to show the irregularity of the more turbulent cells. In the graph, the corresponding cumulative densities of cell area distributions show that more turbulence creates larger and more irregular cells with wider distribution of normalized surface areas.

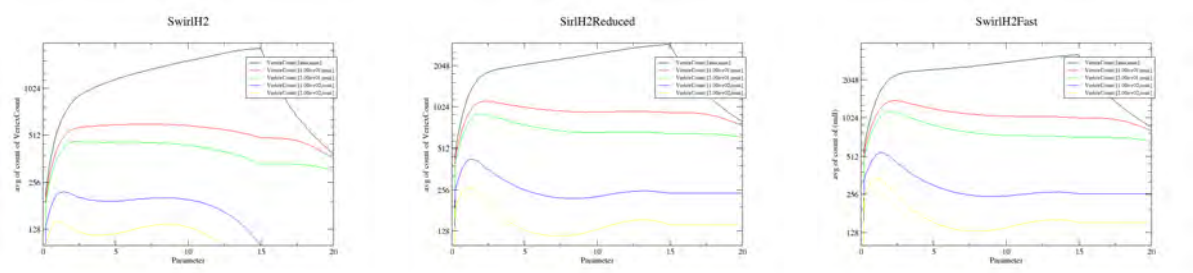
3D segmentation interactively. Based on these subsets, we aggregate pre-computed statistics for quantitative analysis.

Thus, topological analysis allows one to fully explore the parameter space used for segmenting, selecting, and tracking burning cells as defined by the domain scientists. Furthermore, we demonstrate that the high level of abstraction of the topological representation, which reduces data by more than two orders of magnitude, does not impact adversely the functionality in the data exploration process. In particular, one can explore all possible segmentations of burning cells to understand better their dynamics as well as to validate the method. Finally, we demonstrate that the topological data representation is ideally suited for performing extensive data analysis by providing a compact, yet complete representation of features of interest.

We continue to provide new analysis capabilities. For example a recent study of cell counts conditioned on size has revealed interesting new features that we, in collaboration with the stakeholders, are currently working on interpreting. Even a mild conditioning on cell size markedly changes the shape of the plot revealing a large plateau of cell counts in which the cell counts are no longer dependent on threshold, see Figure 6.

The Impact.

Topological techniques allow us to compute all necessary information for a complete segmentation and analysis of burning cells in a single pass. Furthermore, the resulting data is small and can



(a) Number of cells vs fuel consumption threshold for the SwirlH2 data set.

(b) Number of cells vs fuel consumption threshold for the SwirlH2Reduced data set.

(c) Number of cells vs fuel consumption threshold for the SwirlH2Fast data set.

Figure 6: These images show the relationship between cell counts and fuel consumption for different datasets.

be handled efficiently on standard computers. This allows, for the first time extensive parameter studies on the burning cells which previously would have taken months of computing time. In particular, our methods have been used to show that in an idealized H2 flame the average cell size increases with turbulence a somewhat surprising and counter-intuitive result.

2.6 Combustion – Topological Analysis of DNS Simulation Results

The Problem. Detailed simulation of fundamental combustion processes are an important tool to improve engine and power plant design. In particular, simulations are aimed at understanding transient effects difficult to observe experimentally, such as local flame extinction. The goal of this project is to provide scientists with new tools to define and analyze features of interest, e.g. extinction / re-ignition regions, robustly and efficiently in large scale data.

The Solution. We are using topological approaches to define and extract features of interest. These include, for example, the definition of extinction regions as local level sets, the analysis of dissipation elements as crystals in the Morse-Smale complexes, or the extraction of ridges as Jacobi sets. For each topological structure we have developed or are in the process of developing discrete algorithms that due to their combinatorial nature are unaffected by numerical problems and can be implemented robustly independent of the complexity of the underlying data. Furthermore, we take advantage of the inherently hierarchical nature of topological structures to remove noise and analyze the data at different scales. This allows a stable analysis of, for example, dissipation element so far have been considered notoriously unstable and sensitive to noise.

The Impact. Our techniques have enabled the extraction and analysis of a variety of features of interest in large scale turbulent combustion simulations. These have provided new insights into the creation and evolution of extinction / re-ignition regions that are of fundamental interest to the scientists. In particular, we have provided new feature based statistics that go beyond the traditional global characterizations of flame behavior and allow detailed, local analysis.

Progress this Period.

Identification of High Scalar Dissipation Structures Using the Morse-Smale Decomposition.

We focus on data from the JET simulation, a temporally-evolving turbulent CO/H2 jet flame undergoing extinction and reignition at different Reynolds numbers. The simulation was performed with up to 0.5 billion grid points. The configuration is shown on the left of Figure 7. Periodic

boundary conditions in the mean flow (x) direction results in a situation where the mixing rates increase until approximately midway through the simulation, after which point they begin to decay.

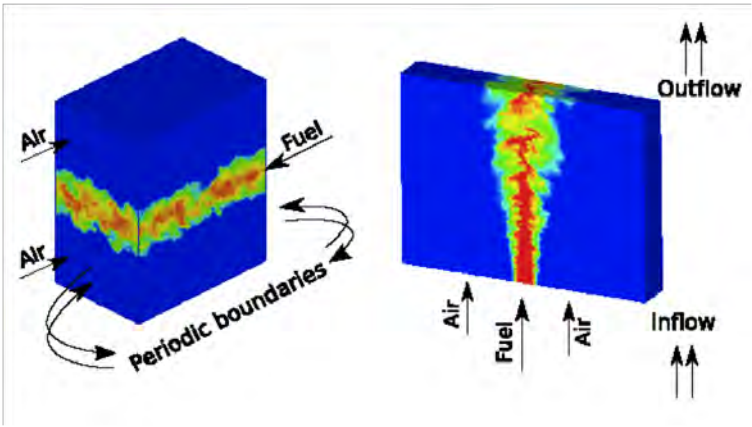


Figure 7: Our team is working with output from the JET simulation (J. Chen, SNL-CA).

We derive a number of relations among the species of the simulation that allow formulating hypotheses of the dynamics of the extinction and re-ignition process. We have computed the Length Scales from statistics of the thickness of the pancake-like scalar dissipation features. Thickness is known to increase over time in JET. However, full characterization of the dissipation feature involves other structures as well, and we compute basins in the mixture fraction field to correlate with the pancake-like feature in scalar dissipation as well as high vortical regions in the velocity field.

We have implemented a new code base for computing the MS complex required for this segmentation, and have produced some initial segmentation and visualization results. The new code has proven to be scalable, and we have produced this analysis at the highest resolution of the data.

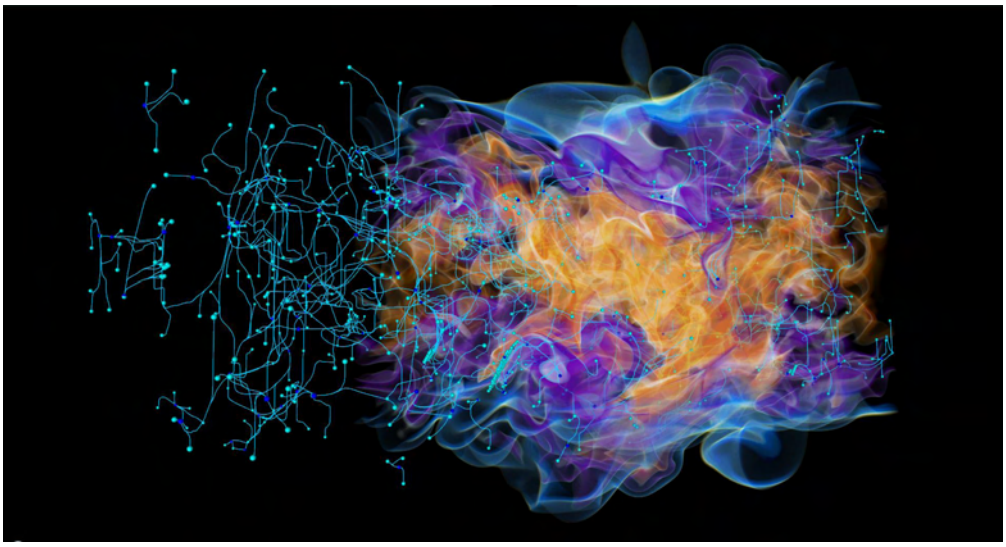


Figure 8: The "spaghetti" structure connecting minima in the mixture fraction represents the connectivity of basins.

In Figure 8, we show the stable minima in the 0.4 to 0.6 range of mixture fraction. These are the centers of the basins of mixture fraction, and we have computed some preliminary statistics characterizing these regions.

Our new code base allows for a full 3D segmentation and production of high-quality renderings, such as in Figure 9. Here, the 3D basins around the minima are rendered.

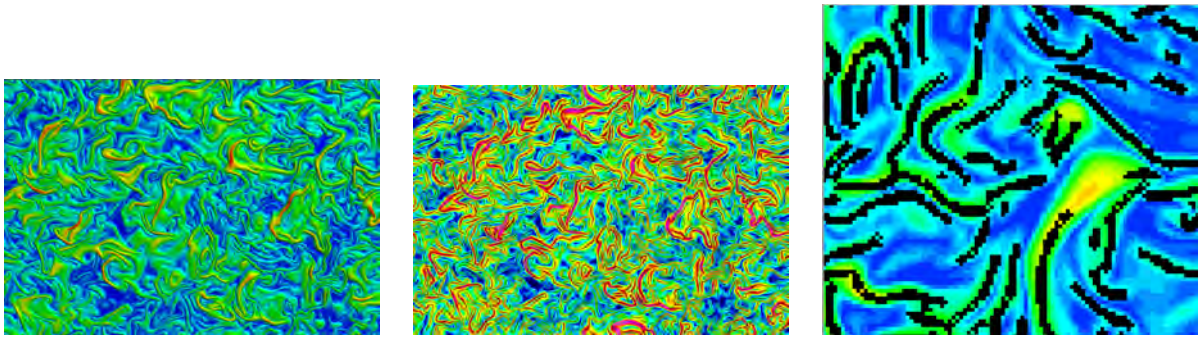
(a) Volume rendering of mixture fraction.

(b) Volume + basins.

(c) Basins extracted.

Figure 9: Visualization showing volume rendering (left), volume rendering along with the basins (middle), and just the basins we compute as part of our analysis (right).

High Scalar Dissipation Structures as Ridge Lines. So far we have segmented high scalar dissipation regions as level sets surrounding maxima in the χ (chi) field. However, for two-dimensional experimental data the traditional method of extracting and analyzing similar structures (using the magnitude of temperature gradient as a stand-in for χ) exist based on computing "ridge-lines." In this context a point is classified as ridge if it is a local maximum in the direction orthogonal to the local gradient. Unfortunately, this criterion is not stable and multiple heuristics are required to achieve an adequate segmentation. Furthermore, initial experiments in 3D using a generalized definition of a ridge have failed to produce useful segmentations (Figure 10). Nevertheless, comparing topological techniques with traditional methods is an important step to validate the new methods. Additionally, replicating ridge like structures in two-dimensional data using topological techniques may provide important insights into the unsolved problem of computing "ridge-surfaces" in the volumetric case.



(a) χ field at y=448 of time step 100 drawn using a logarithmic color map.

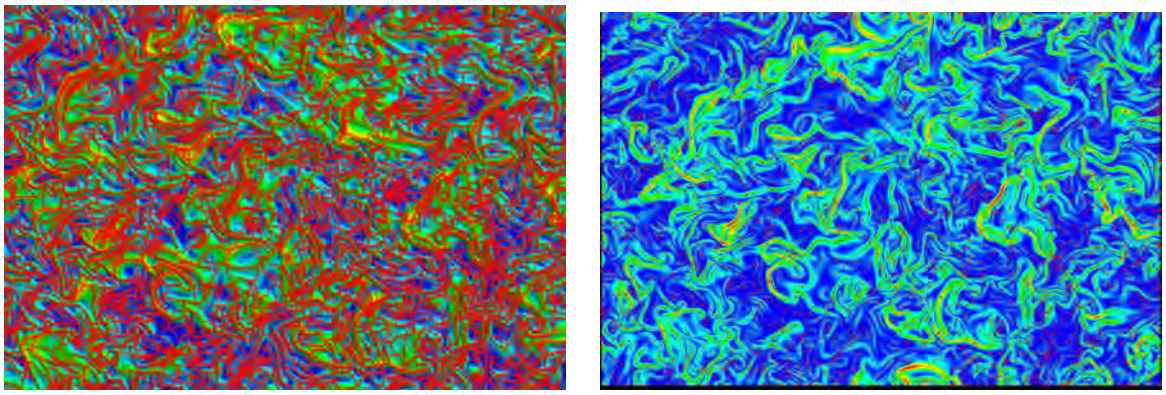
(b) Ridge lines in the χ field as extracted by traditional methods.

(c) Zoom-in to the ridge line segmentation showing ridges are multiple pixels thick and can become degenerate.

Figure 10: The traditional approach for computing ridge lines is not stable and fails to produce useful segmentations.

One of the main problems of traditional ridge-line extraction is that it is an entirely local process. Each pixel is classified independently leading to numerical instabilities and structural inconsistencies. For example, the close-up on the ridge-lines (Figure 10(c)) shows that ridges are multiple pixel wide and can degenerate into ridge regions which violates the definition as well as the intuition of a "ridge-line." At the same time, the color map suggests that ridges are related to

ascending lines of the MS complex. While there does not yet exist a mathematical theory linking both structures, experiments seem to support this notion.



(a) All ascending lines of the MS complex of the χ field shown above.

(b) Ascending lines of the MS complex filtered by gradient magnitude.

Figure 11: Computation of the MS complex and its ascending lines from ridge lines using the traditional method.

Initially, the ridge lines seem to be a subset of the ascending lines in the MS complex, and simply filtering all unnecessary lines would provide a segmentation highly similar to the traditional approach. By definition, the ascending manifolds cannot degenerate and they are computed in a global approach making them far less likely to be influenced by numerical instabilities in the data. Furthermore, there exists a direct generalization of ascending line for the volumetric case where the ridge surfaces should be described by the two-dimensional ascending manifolds in the χ field. However, as Figure 12 illustrates, upon closer inspection, not every "ridge" is identified as an ascending manifold of the MS complex.

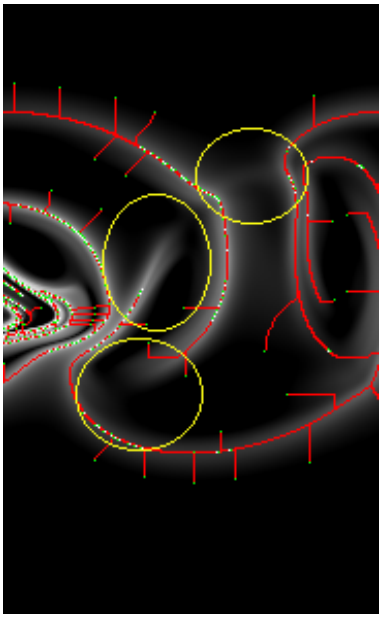
The fundamental problem with using the MS complex in a straightforward manner is that "ridges" are more intuitively defined on local curvature than critical point and gradient behavior. In particular, several ridges can merge together, and there will not be a saddle at the merge points to start an ascending 1-manifold for each. We investigated a technique for inserting additional saddles by first considering the level set curvature of the "zero" value. Maxima on this curvature function are identified as saddles from which to trace ascending paths. Figure 13 illustrates that while this approach identifies many more of the missing ridges, some are still missing. Furthermore, spurious ridges, called "feathers," are identified that are due completely to noise in the curvature function (Figures 12 and 13).

The feathers prove to be extremely difficult to remove. We define "ridgy-ness" criteria based on both persistence for topological simplification as well as various heuristics based on local curvature magnitude. Figure 14 shows how these heuristics were unable to remove certain undesirable features.

The missing ridges lines are due to the fact that we picked the zero level set as an arbitrary threshold for inserting extra saddles. Ridges that merge before reaching the zero level set will not be identified. In fact, if we generalize the notion of where to place this threshold (where to insert extra saddles to guarantee that every ridge will have an arc of the complex), then we converge on the following definition of ridge-lines: A ridge is the Jacobi set of level set curvature and the scalar function. Figure 15 shows that the Jacobi set computed for these two functions in fact identifies every single ridge.

(a) An FTLE field modified to guarantee sufficient saddles to trace each ridge.

(b) Some ridges are still missing.



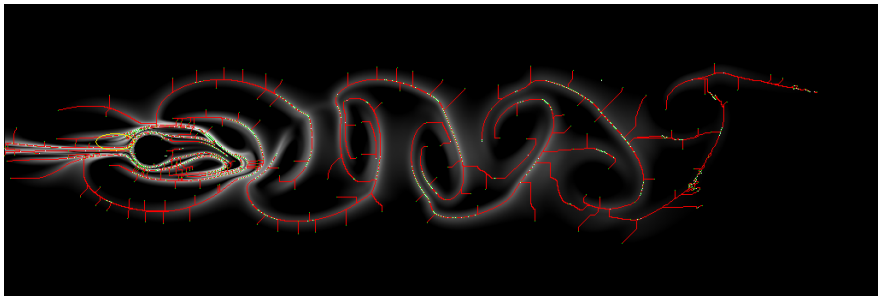
(c) Extra ridges, called "feathers," are difficult to remove.

Figure 12: Compared to the traditional technique, using ascending manifolds cannot degenerate and are computed in a global fashion, making the results less susceptible to numerical "noise" in the data. However, not ever ridge is identified as an ascending manifold using this approach.

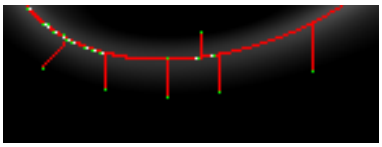
Multi-resolution Ridge Computation.

We are developing a multi-resolution representation of Jacobi sets to be able to capture the ridge structure of the DNS data at the proper scale. The work involves simplification of the ridge structure through concurrent smoothing of the underlying function and of its derivative. The challenge is

(a) An FTLE field modified to guarantee sufficient saddles to trace each ridge.



(b) Some ridges are still missing.



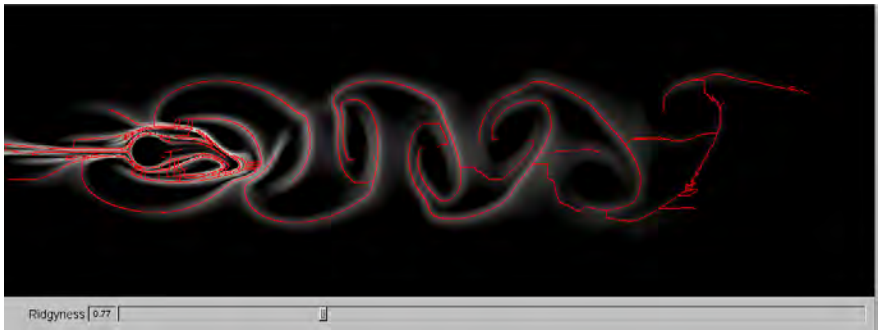
(c) Extra ridges, called "feathers," are difficult to remove.

Figure 13: Exploring an alternative technique for solving the problem of missing and incomplete ridges solves the problem missing ridges, but introduces new, spurious "feather" ridges.

in maintaining the combinatorial structure of the field while performing the simplification. The current preliminary results are encouraging but more research is needed to complete the work and before it can be applied to experimental data.

Figure 16 shows such a simplification prototype. We are currently investigating how to simulate this smoothing in a robust manner that respects the original geometric locations of the ridge lines.

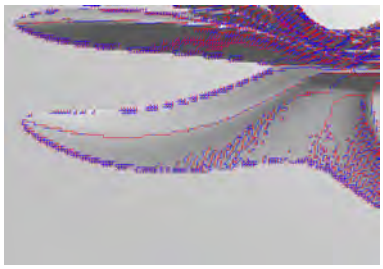
(a) Low threshold.



(b) High threshold.

Figure 14: Even the addition of heuristics is not sufficient to remove all spurious edges. These images show different "ridgy-ness" thresholds for identifying which arcs of the MS complex are ridges.

(a) Using a Jacobi set definition of ridges, we do not omit any ridges.



(b) However, this technique is susceptible to noise.

Figure 15: An alternative definition of ridge lines as the Jacobi set of level set curvature and the scalar function produces a robust technique for finding ridge lines.

2.7 Nuclear Energy

The Problem. To meet the visualization and analysis needs of the NEAMS (the Office of Nuclear Energy's supercomputing program: Nuclear Energy Advanced Modeling and Simulation) and specifically Argonne's Nek5000 code team. The lead of the Nek code, Paul Fischer, is funded by the Office of Science's base program for math and is an INCITE awardee. The needs are met by: (i) providing them tools for bread-and-butter functionality, (ii) providing support, (iii) helping with high end movies, and (iv) helping with high-end analysis.

The Solution. To provide tools for bread-and-butter functionality, we are deploying VisIt to the Nek team. Providing support means we are responding to questions that come from Nek. We

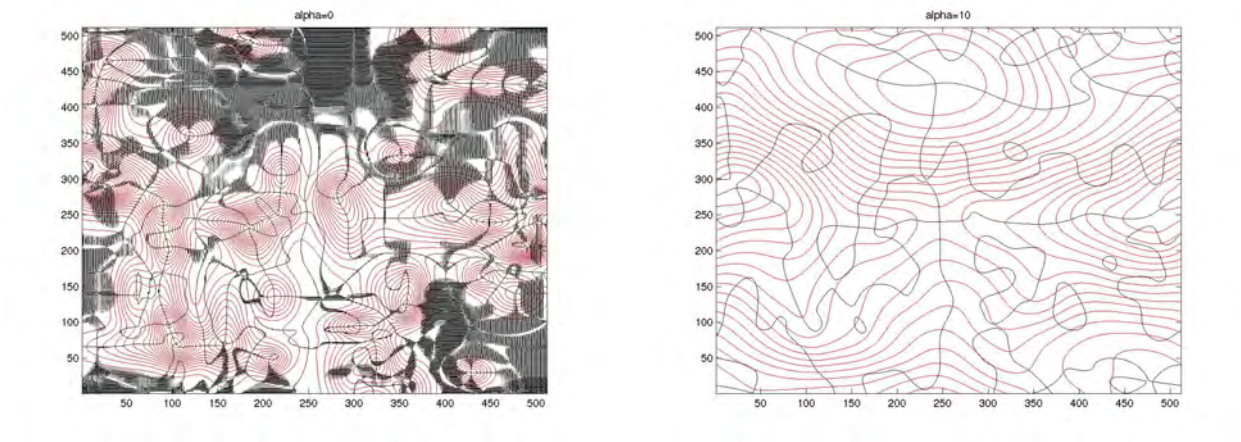


Figure 16: Concurrent simplification of Jacobi set, its function and derivative. The movement of the structure during the simplification shows how the simple removal of edges is an insufficient process to achieve the multiresolution representation desired.

help to create advanced movies and animations for the Nek team. And we help them to perform high-end analysis (examples below).

The Impact. We have met their bread-and-butter needs and have also helped answer some science questions.

Recent Accomplishments.

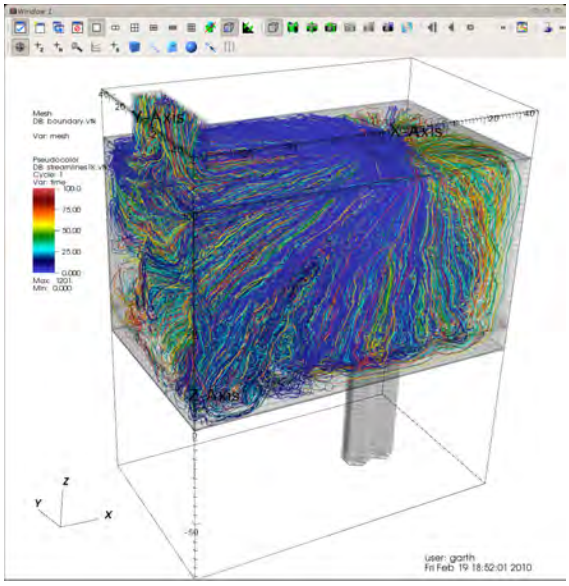
Paul Fischer and Aleks Obabko have run a simulation of air flow in a box (the "fish tank"). The purpose of these simulations are code validation. Argonne has invested institutional money in building a facility to reproduce this simulation. The simulation, in turn, was used to tune the experiment. The experiment is upcoming, where velocities will be observed on a side of the fish tank box and we anticipate being involved in this analysis.

For the current analysis, Paul and Aleks wanted to better understand the "residence time" of the air. In the experiment, air is pumped in through two inlets (cold and hot air). It exits through a single outlet. They want to better understand how long air stays inside the fish tank before exiting. (Does it circulate for a long time? Does it exit immediately?) We are able to answer this question by placing particles in the inlets and observing their behavior.

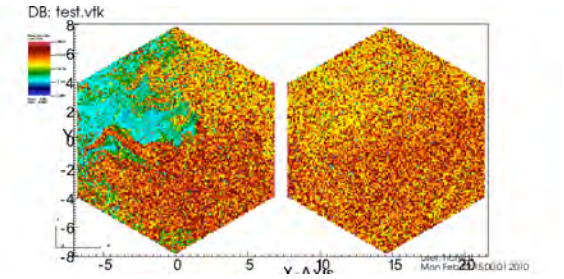
2.8 Computational Chemistry and the MADNESS Team

The Problem. The MADNESS code currently lacks almost any sort of scientific visualization and analysis postprocessing capability. There is not one simple scientific goal, but instead we are attempting to enable scientific investigation, introspection, code debugging, and data exploration.

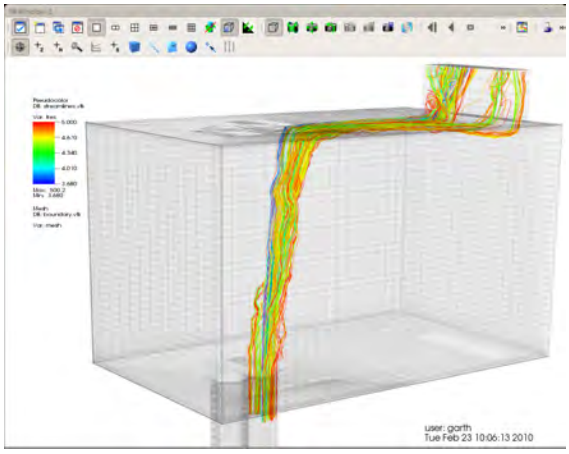
The Solution. We plan to deliver these capabilities by utilizing as much of VisIt as possible, at least initially enabling the scientists to explore their data directly. For other projects this could have been as simple as writing a new file format reader. Here, however, we lack a data file format entirely, and the data model is a poor match for the VTK architecture on which VisIt (and similar scientific tools) are based, as it is a deep quadtree-AMR structure with extremely high-order elements (up to 20th order) and up to a 6 dimensional basis. Therefore, the data file format is being developed as a collaboration between the MADNESS developers and the VACET team, and file format readers and support infrastructure for VisIt written with the special requirements of their data model in mind.



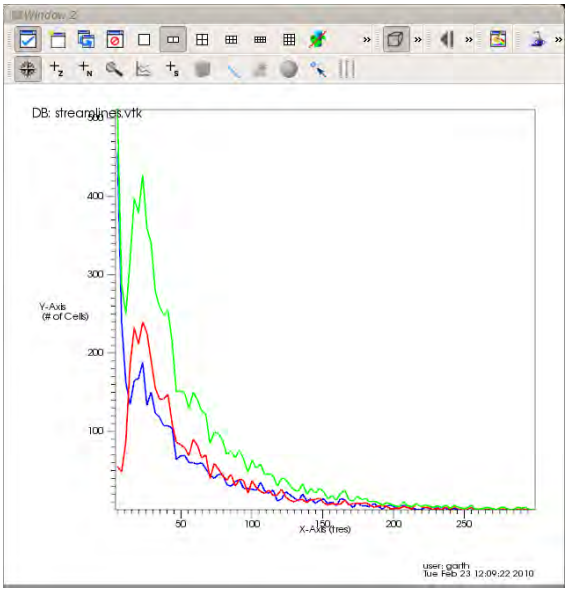
(a) An illustration of 10,000 streamlines moving through the fish tank.



(b) We placed particles at regular intervals along the inlet. We then advected the particles and calculated the residence time. This picture colors the initial position of the particle by the duration of residence. You can see a "fast track" on the left inlet, where particles are able to get to the outlet more quickly.



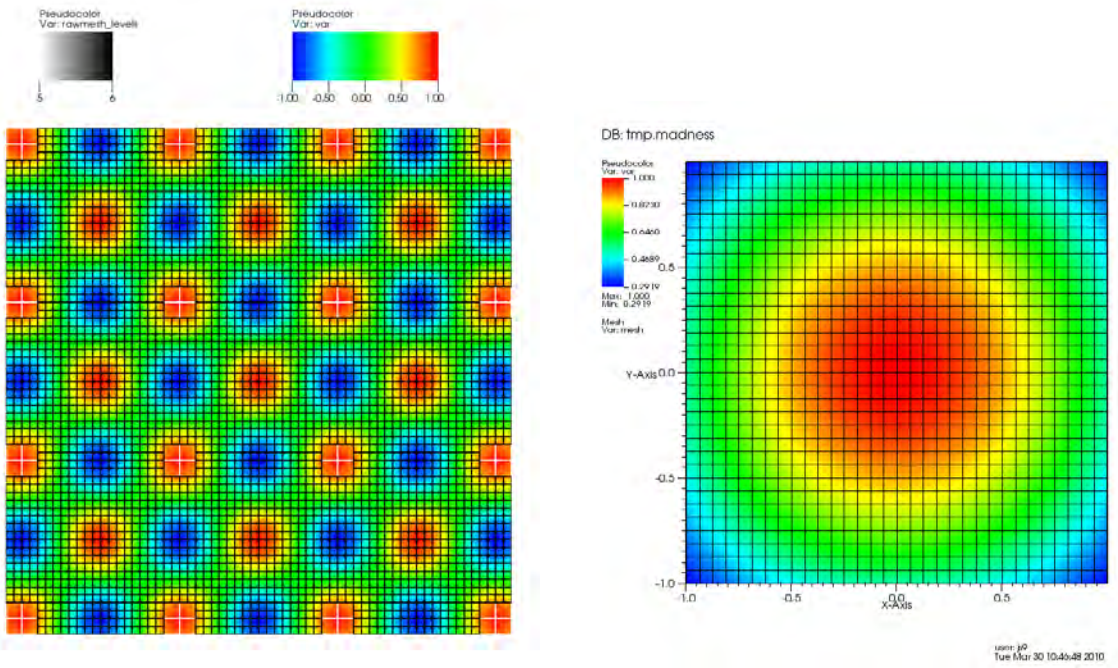
(c) We rendered only those particles with quick residence time. These particles correspond to the cyan areas of the previous image.



(d) A histogram of residence time (the original science goal). The blue line is the cold air inlet, red is host, green is combined.

Figure 17: This sequence shows different stages of an analysis of the "residence time" of air in the "fishtank" experiment.

The Impact. By pursuing this project, we will enable many types of scientific discovery using the MADNESS simulation code which are difficult, or impossible, without robust analysis and visualization tools. Specific scientific problems are not yet being addressed, as we are not yet at the



(a) Visualization of plain-text dump of MADNESS structures.

(b) Resampling of the MADNESS quad-tree hierarchy onto a regular mesh.

Figure 18: Example progress of work-to-date with the MADNESS team.

stage where analysis can even be performed, let alone identifying shortcomings with the available tools.

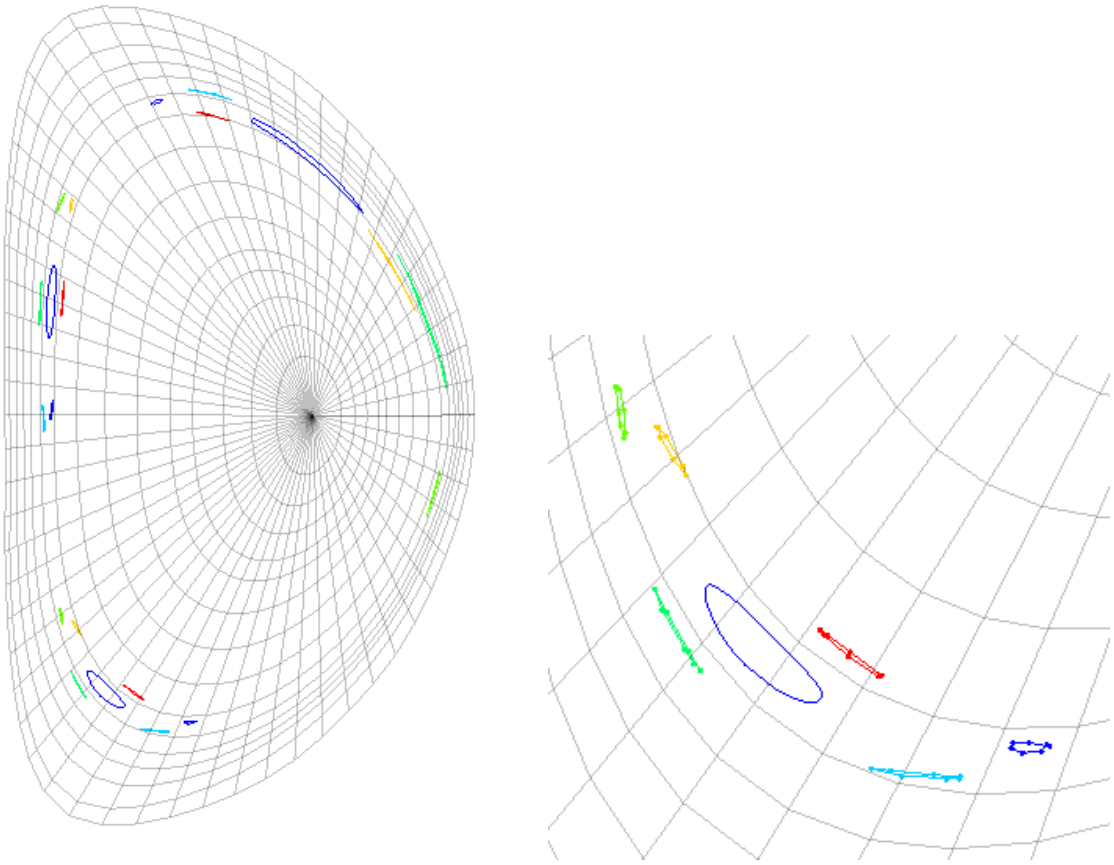
2.9 Fusion: Fieldline Analysis and Poincaré Plots

The Problem. Physicists are currently studying the affects of magnetic islands that form in plasma. These islands cause defects in the magnetic field and the current flow resulting in contact between previously separate regions. This contact results in hot areas coming into contact with cool areas which leads to core cooling. A common way of viewing simulation results with an eye towards understanding island formation is the *Poincaré plot.* Physicists would like to have tools that allow them to be able to automatically generate Poincaré maps of the magnetic field and detect the island formation and track them over time. To help increase understanding of islands, it is helpful to display other simulation variables in conjunction with the Poincaré plot. Such data may be scalar (electric potentials) or vector data (magnetic fieldlines) and may have its own visualization requirements.

The Solution. Our objective is to develop tools for the robust analysis of toroidal fieldlines, include high quality Poincaré plots, and to deploy these new technologies in production form to the fusion community.

The Impact. These new visualization techniques will help to increase scientific understanding about heat transport in highly stochastic fields, such as those produced by DOE fusion simulation codes. We are deploying these new capabilities to multiple fusion code teams in DOE to achieve wide impact. Our deployment vehicle is VisIt, a production-quality visualization application. Thus, our R&D effort is able to leverage a large, existing investment in software and release engineering.

Accomplishments This Period.



(a) Poincaré Plot showing a "3,1 island chain" that is composed of six islands within itself – islands within islands. Seeing such fine structures is important for understanding heat transport in highly stochastic fields.

(b) A closeup of one set of islands within islands. For reference, a sibling island chain (blue) is nested within the six islands-within-islands.

Figure 19:

The major emphasis has been on the release of VisIt 1.12.2. This release is the first in which the new Poincaré tools are fully functional. The R&D work to achieve this objective greatly enhances the analysis capability for the fusion research community to the point that we are now able to focus on other needs that lay more on the research side (as opposed to infrastructure).

For instance, we successfully deployed a M3D C1 reader and integrator into VisIt that allowed for using higher order elements. This deployment allows for both the calculation of the magnetic fieldlines as well as the visualization of scalar data. Previously, the M3D C1 researchers could not view fieldlines and had to generate their plots in a batch mode. Further, for viewing scalar data they have had to rely on less robust tools. Our deployment has beneficial in that we were able to detect bugs in the their simulation that previously they were unaware of.

At this point in time we have successfully used the tools on output from four different simulation codes, NIMROD, M3D, M3D-C1, and Siesta.

We have had a few setbacks in the 2.0 VisIt update in that we have discovered that features such as some parallel computations do not work as expected, as well as problems with the VTK infrastructure. These bugs are all fixable. However, utilizing VTK for the calculation of streamlines

in VisIt is overhead intensive (5x times slower than in SCIRun). As such, the decision has been made to instead focus on using new computational technology emerging from the VACET integral curve effort (Section 3.3).

2.10 Fusion: Query-Driven Visual Data Exploration and Analysis

The Problem. Many fusion simulation codes use the particle-in-cell method, which produces data files containing millions or billions of particles per time step. The sheer size and complexity of such data is an impediment to scientific understanding.

The Solution. We are leveraging work in query-driven visualization, such as that we performed previously for an accelerator modeling project, to enable rapid (interactive) visual exploration of large, time-varying particle-based datasets. We are extending previous techniques, which include a combination of novel user interface and index/query technology[2], for use on fusion applications. The ultimate objective is to provide specific features needed for the fusion community, features that go above and beyond those that were sufficient for previous projects with the accelerator community.

The Impact. Once completed and deployed, our target is to help improve scientific understanding of the radial transport mechanisms in gyrokinetic simulation codes. This objective will help to enable progress in fusion science.

Accomplishments this Period.

The high level objective/milestone for this project is a robust system for query-driven visual data exploration and analysis of particle-based fusion data. The system should allow the application scientist to display their data using parallel coordinate then form a "cumulative query" and display the results.

The major emphasis this period has been on the release of VisIt 1.12.2. This release was the first in which the query based tools were fully functional. However, there is still much hardening of the tools that needs to be done. We anticipate that as scientists use the tools, we will find areas that need work. For instance, the H5Part/HDF_UC Reader was found to have multiple shortcomings in terms of creating and reading the FastBit indices; isolating those shortcomings required a significant effort. We implemented bug fixes and workarounds, including support code for creating FastBit indices that does not rely on the HDF_UC library. This code was used to convert particle files into the H5Part format; multiple conversion programs have been deployed to GTC and XGC users.

The continued bottleneck to having the tools used by the fusion community is the lack of a cumulative query option. This work was to be the main focus of the Fusion SAP but has stalled because the other infrastructure to support queries in general has been overwhelming.

2.11 Advanced Visualization in the SDM Dashboard

The Problem. Simulations that require massive amounts of computing power and generate tens of terabytes of data are now part of the daily lives of scientists. Analyzing and visualizing the results of these simulations as they are computed can lead not only to early insights but also to useful knowledge that can be provided as feedback to the simulation, avoiding unnecessary use of computing power. Our work is aimed at making advanced visualization tools available to scientists in a user-friendly, web-based environment where they can be accessed anytime, from anywhere.

The Solution. We explore a web-based analysis and visualization solution to this problem in the context of turbulent combustion simulations. To understand the coupling between turbulence

[2]We are using FastBit in collaboration with the SciDAC Scientific Data Management Center.

and the turbulent mixing of scalars, such as temperature and species concentrations, it is important to generate isosurfaces that represent those interactions.

Isosurfaces are one of the most widely-used visualization techniques and efficient to compute: the complexity of standard marching cubes, the most popular isosurface algorithm, is linear. Although it is possible to efficiently generate an isosurface for a given isovalue, computing and rendering a large number of isosurfaces, as required in this scenario, is expensive and incurs a high network overhead for transferring the results to a web browser. This makes such a solution impractical for a web-based analysis tool. To address this problem, we propose the use of a summary structure, called contour tree, that captures the topological structure of a scalar field and guides the user at identifying useful (important) isosurfaces.

We have also designed a user interface that allows users to interact with and effectively explore multiple isosurfaces.

The Impact. By applying the contour tree algorithm to find isosurface values in-situ with the computation, it is possible to selectively browse through multiple visualizations and quickly understand the complex data being generated during the simulation. The contour tree tool has been integrated with the eSimMon dashboard system, which provides an environment for scientists to monitor, manage and explore simulation results. We have done a short case study where we show that integrating the dashboard with the interactive contour tree tool leads to an effective and efficient means to explore the turbulent combustion simulation results.

Future Work. Future plans will focus on integrating VisIt into the SDM dashboard.

3 Technology Incubation Projects

3.1 Hybrid-Parallelism and Volume Rendering

The problem. Modern computational platforms are evolving towards using multi-core processors; future generations of machines will be built using processors containing tens or hundreds of cores. There is concern that existing, traditional message-based parallel programming models will not scale well on such platforms. The aim here is to better understand how well hybrid-parallelism, which combines both traditional message-based distributed memory parallel concepts with multi-core, shared-memory parallelism, performs for visualization algorithms, raycasting volume rendering specifically, as compared to traditional message-based, distributed-memory parallelism.

The solution. Our solution entails conducting performance and scalability tests of traditional and hybrid parallel implementations of raycasting volume rendering, a staple visualization algorithm. Our approach is to compare performance using several different metrics: (1) absolute runtime, (2) memory footprint at various stages of algorithm execution; (3) communication characteristics of the two implementations.

Impact. The results of this study show that the hybrid-parallel approach offers clear and distinct performance advantages when compared to the traditional approach to parallelism. First, the hybrid-parallel version consumes between one sixth and one twelfth the amount of memory required by the traditional MPI version just for initialization. Second, at high levels of concurrency, the hybrid-parallel implementation runs about three times faster. Third, the hybrid-parallel version requires only about half the communication bandwidth compared to the MPI version. These early results will help to shape the architecture of future visualization and analysis applications so as to be able to run effectively on current petascale and future exascale platforms.

Resources. For these runs, we ran strong scaling studies on Franklin at NERSC and on JaguarPF at ORNL. Our time at NERSC comes through the yearly ERCAP allocation process,

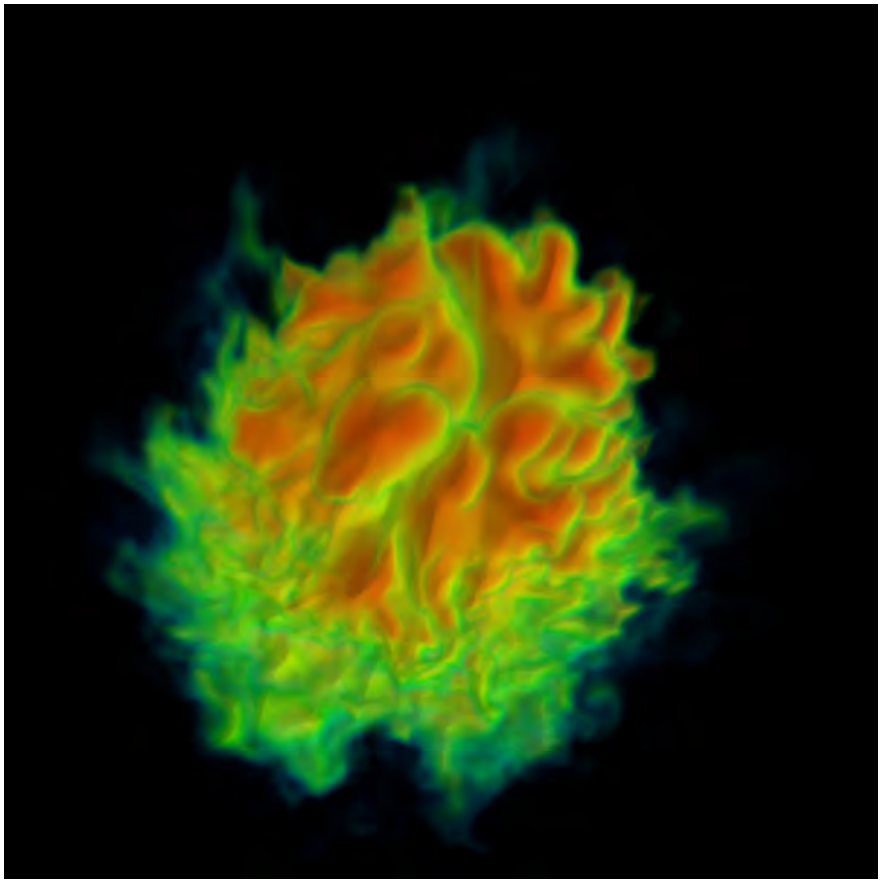


Figure 20: Example output from the hybrid-parallel volume rendering application, input data is results from a combustion simulation of a hydrogen flame. The source data was about 4Kx4Kx4K and the application run at 216K-way parallel on JaguarPF at ORNL.

while our time at JaguarPF comes through a Director's Discretionary allocation to S. Ahern for visualization work.

Upcoming Plans. Weak scaling study. Our early work on this project, which resulted in a publication in the Eurographics 2010 Parallel Graphics and Visualization Symposium, showed scalability characteristics up to 216K-way parallel using a strong scaling study. Our future plans will focus on studying scalability characteristics at these same levels of extreme concurrency but using a weak scaling study.

3.2 Hybrid-Parallelism and Streamlines

The problem. Integral curve techniques are used in various visualization and analysis activities by VACET. Our basic parallel algorithm is the one described in the SC09 paper by Pugmire, Childs, Garth, Weber, and Ahern. However, we feel there is an opportunity for improved performance in a hybrid parallel setting. As a secondary goal, we wish to gain experience in the hybrid parallel space with production visualization tools (VisIt/VTK). We wish to better understand the barriers in deploying a production visualization tool in a hybrid parallel setting. This is highly relevant to the Office of Science mission, since highly multi-core nodes will be part of the Office's future supercomputers.

The solution. The SC09 paper had three basic parallelization techniques: parallelize over

(a) Magnetic field computed by a NIMROD simulation.

(b) Fluid flow in a NEK3D thermal hydraulics simulation.

(c) Magnetic field in a GenASiS supernova simulation

Figure 21: Example images of streamlines computed with our hybrid-parallel method.

data, parallelize over seed points, and a hybrid that parallelized over both data and seeds ("master/slave"). We adapted the first two techniques to work in a hybrid parallel setting. We expected many potential benefits.

For hybrid parallelization with parallelizing over data:

1. Load imbalance is the key issue with this technique. By using hybrid parallelism, a larger number of cores can access an MPI task's portion of the data set. This will decrease load imbalance.
2. Communication is a significant amount of the overall execution time. With hybrid parallelism, there are fewer cores participating in the communication, improving performance.

For hybrid parallelizations with parallelizing over seeds:

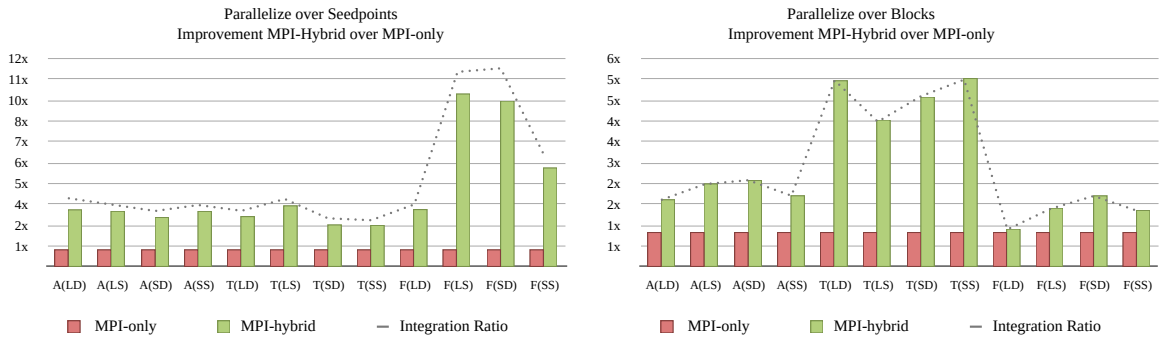
1. We expect reduced I/O, since blocks can be shared among cores on the same node.
2. We expect reduced I/O, since the cache size on a node is larger.
3. We expect better parallel efficiency. In a non-hybrid setting, some MPI task is given seeds that are "harder", whether that means traversing more blocks, integrating for longer, etc. In a hybrid setting, each MPI task receives more seeds, but has more cores to operate on those seeds. The net effect is to average out extreme behavior, which is important because the algorithm takes as long as the slowest seed.

In our study, we reproduced the tests in the SC09 paper, looking at multiple data sets and seed point configurations. We saw benefits of up to 8X performance improvement. Since we were running 4 threads per node (franklin), we expected at most 4X improvement (Figure 22. The additional performance was possible only because of the combination of the factors listed above.

Impact. We have developed a greater understanding of the benefits and pitfalls of hybrid parallelism. We established that hybrid parallelism significantly improves streamline performance. We also learned about a variety of pragmatic issues to deploying a production visualization tool in a hybrid environment.

3.3 Integral Curves: Streamlines and Stream Surfaces

The Problem. Modern simulation codes increasingly involve the simulation of vector fields that appear in problems relating to astrophysics, thermal hydraulics, fusion, and fluid flow. Existing



(a) Parallelization over streamlines (load-on-demand).

(b) Parallelization over domains (parallel static domains).

Figure 22: Performance comparison of two different hybrid-parallel approaches with a traditional MPI-only implementation. The hybrid-parallel approach shows consistent performance advantages over the traditional approach over twelve different sample problems.

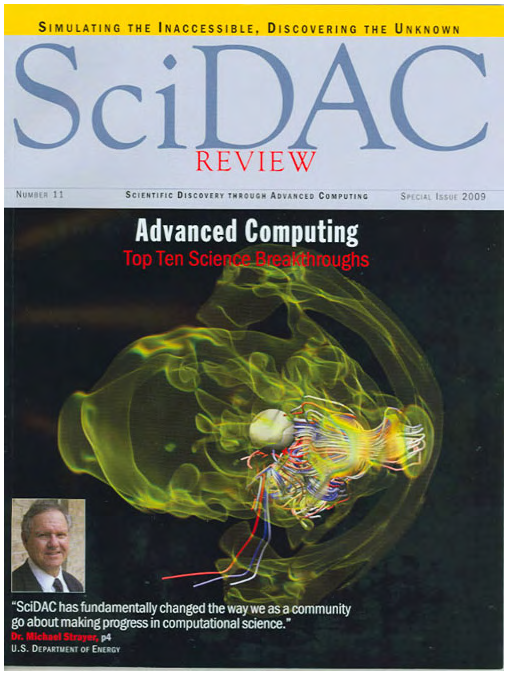
visualization and analysis capabilities for scalar fields cannot fulfill the need for adequate visualization of such vector fields, as the delocalized nature of typical phenomena described by vector fields such as e.g. transport and mixing is not adequately captured by scalar-based approaches. For such problems, modern integration-based methods are used that derive visualization and analysis from the trajectories of idealized massless particles. These methods provide much increased insight into vector fields, but are computationally intensive and challenging in their application to very large (petascale) datasets. The aim of this project is to enable physicists, chemists and fluid dynamicists to visualize and analyze their state-of-the-art simulation data using robust, efficient and scalable integration-based visualization techniques over a wide variety of data representations and problem characteristics.

The Solution. Our solution is based on a novel code framework, integrated into VisIt, that enables the efficient and scalable computation of integral curves in vector fields represented over regular, structured, unstructured and AMR meshes. Due to the large variation in problem characteristics involving integration-based visualization algorithms, our framework can leverage several distinct parallelization schemes that are uniquely suited to specific problem parameters. Also, we provide an adaptive scheme that allows scientists to conduct vector field visualization without having to familiarize themselves with parallelization strategies, allowing for rapid adoption of integration-based methods.

To increase performance and scalable efficiency, two of the parallelization schemes make use of hybrid parallelism, where fewer MPI tasks are used (typically one per node) and shared-memory parallelism is employed within a node. This approach, although more challenging to implement, can enable significant performance and efficiency gains on supercomputers based multi-core CPUs (Section 3.2).

Furthermore, building on these capabilities, we are investigating novel integration-based visualization algorithms, such as integral surfaces and Lagrangian techniques, that can provide adequate abstraction for specific vector field phenomena. The aim of this project to deliver a comprehensive solution for integration-based visualization based on both baseline integration capabilities as well as advanced visualization algorithms to science stakeholders in the VisIt visualization tool.

Impact. Based on our framework, scientists will be able to leverage modern integration-based visualization techniques to visualize and analyze vector field data to investigate phenomena such as transport and mixing. Our work specifically addresses very large / petascale data, to enable



(a) VACET's streamline algorithm produces an image showing fluid velocity from a Type II core collapse supernova simulation. (Cover of SciDAC Review, Special Issue, 2009.)

(b) New visualization techniques developed in VACET produce insight into the complex structures of vector fields (Cover of SciDAC Review, Winter 2009.)

Figure 23: VACET's new integral curve algorithm, which has been deployed to the community in VisIt, produces images that appear on two recent issues of SciDAC Review.

robust analysis on current and future-generation datasets (Figure 23). Furthermore, we will provide necessary infrastructure and tools to build improved visualization tools that address the need to analyze simulations of ever-growing complexity.

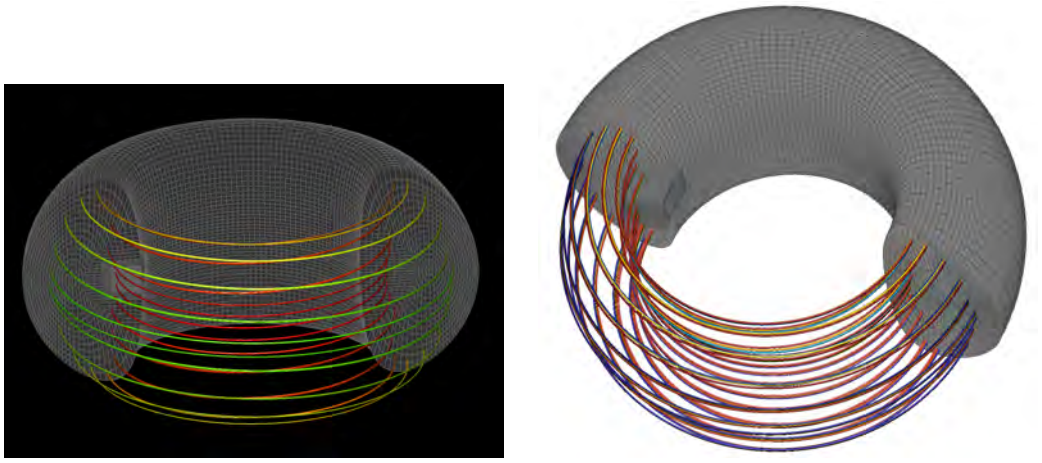
Upcoming work. In the near term, we are focusing on performance evaluation and improvement, which is a requirement for our fusion science stakeholders: the integral curve algorithm plays a fundamental role in the analysis of fusion simulation data. Other near-term work will focus on deploying the hybrid-parallel software (Section 3.2) into the main VisIt code trunk for production deployment.

3.4 AMR Streamlines

The Problem. Adaptive Mesh Refinement (AMR) is a highly effective discretization method for a variety of physical simulation problems and has been applied to the study of vector fields in different application areas.Integral curves, such as streamlines, streaklines, pathlines, and timelines, are an essential tool in the analysis of vector field structures, offering straightforward and intuitive interpretation of visualization results. For integral curve visualization in AMR data, techniques considering the AMR structure are necessary since traditional, unigrid approaches for solving the integral curve problem are ineffective in dealing with AMR datasets.

The Solution. We investigated existing streamline construction algorithms and apply them to AMR vector fields. The goal of this effort was to identify problems that preclude efficient and accurate visualization of AMR vector fields. We have developed a novel algorithm to address these problems. We have performed a numerical evaluation of the accuracy of our method, compared it

against existing approaches, and document the applicability of our scheme by using it to visualize AMR data sets from astrophysics, flow, and fusion simulations (Figure 24).



(a) Multiple streamlines in an AMR mapped grid fusion dataset.

(b) Single (long) streamline in an AMR mapped grid fusion dataset.

Figure 24: Examples of our new AMR-aware integral curve algorithm applied to mapped-grid AMR datasets from our fusion stakeholders.

The Impact. The new technique provides feasibility to visually analyze vector field data in application areas, where the simulation must accommodate many different spatial scales. For example, in simulations of the solar system, the computational domain may span thousands of astronomical units (AU), whereas some physical structures that need to be resolved occupy significantly shorter length scales.

Upcoming Plans. Near-term objectives will focus on a combination of continuing to refine/debug the new technique with scientific data (fusion simulation results on mapped AMR grids, specifically), and integrating the new techniques into the VisIt code trunk for widespread deployment to the scientific community.

3.5 Multiple-GPU Volume Rendering

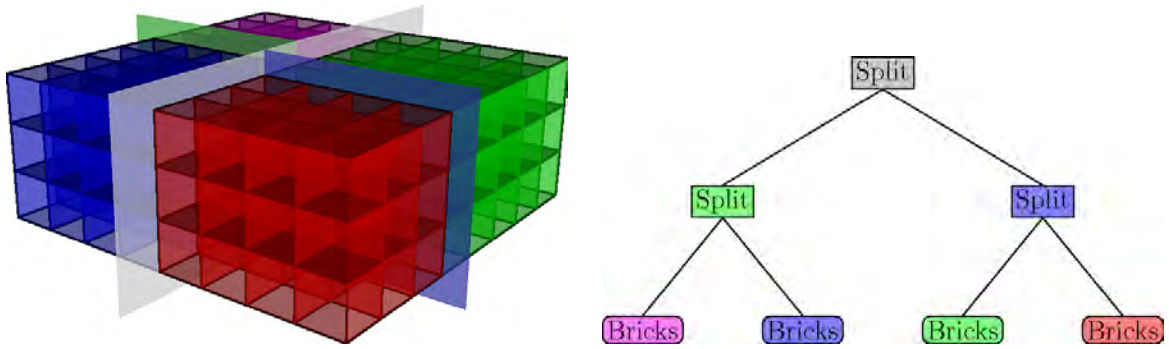
The Problem. SciDAC scientists need to understand complex phenomena at the small and large scale. One of the ways scientists do this is by visualizing their data through a method dubbed "volume rendering." As data sizes grow, visualization and analysis processes, including volume rendering, must scale with the data. Otherwise we risk wasting the large-scale results needed to resolve features of a simulation due to an inability to gain insight from that additional resolution.

The Solution. Volume rendering algorithms can be broadly classified as those implemented to utilize CPUs versus those implemented to utilize graphics processing units (GPUs). Typical parallel volume renderers use CPUs due to conventional hardware abilities on large-scale supercomputing resources, yet GPUs have a significantly better price/performance ratio. However, in recent years many DOE funded sites, such as TACC, LLNS computing, and NCCS have started acquiring GPU-accelerated parallel computing resources. We have performed research and development to run GPU-accelerated volume rendering on these newer computing resources, achieving multi-scale parallelized methods which are more efficient than CPU-based implementations, yet utilize a fraction of computational resources.

The Impact. This project is still in the incubation and research stages. When completed, it

will allow SciDAC researchers to visualize and understand the data sets being generated on tera- or even peta-scale resources.

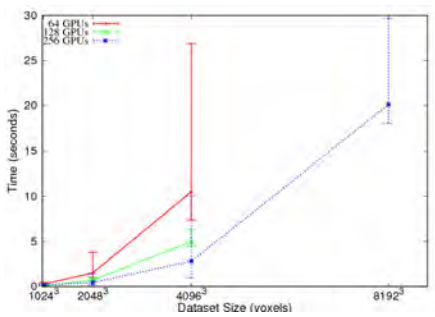
Accomplishments this Period



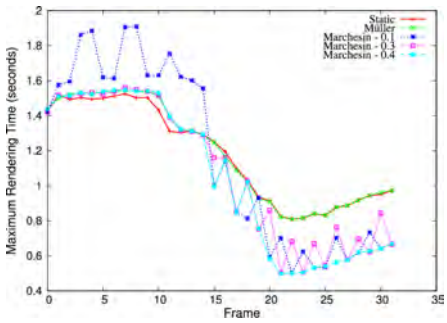
(a) Domain partitioning colored by processor.

(b) kD-tree corresponding to the bricks at left.

Figure 25: Strategy used for packing bricks in parallel volume rendering work. Groups of bricks end up on a single processor; the splitting plane and camera location gives a compositing order.



(a) Tuvok scaling on Longhorn up to data sets of size 8192^3 and 256 GPUs. Data sets that are 2048^3 or smaller can be rendered interactively.



(b) Maximum rendering time across all nodes, per frame, for various load balancing algorithms we implemented in VisIt. Load balancing is a very difficult problem, and we don't recommend it for production vis tools.

(c) Example visualization produced from the scaling study.

Figure 26: Scalability and rendering performance measurements.

- **Assessment.** We wrote and submitted a paper to High Performance Graphics detailing the scalability of the system (Figure 26(a)). For very large data, the primary bottleneck is rendering time; we should focus on improving the performance of the volume renderer.
- **Clusters.** Various fixes available in the "research code" now need to be ported to "production."

- **Tuvok.** Was released as part of ImageVis3D. In Visit, Tuvok is working well as "research code." Older version integrated into VisIt, but in a bad state; needs some engineering time/effort.
- The Tuvok volume renderer won a DoE OASCR at SciDAC 2009.
- Scaled to data sets of size 8192^3, which may be among the largest sizes every published in open literature (Figure 26(b)).
- Starting Equalizer evaluation, with Rob Sisneros (ORNL).
- Spring 2010 VACET AHM: (1) Discussed extending volume rendering approach to AMR data, to directly benefit stakeholder (APDEC). (2) Discussed multi-GPU parallel acceleration of integral curve calculation. (3) Discussed Mac deployment issues: VisIt in MacPorts. (4) Presented / Discussed / Planned how to effectively make use of multiresolution data in VisIt.

3.6 Uncertainty Visualization

The Problem. As data becomes increasingly large and complex, visualization and data analysis techniques are required that not only address issues of large scale data, but also allow scientists to better understand the processes that produce the data and the nuances of the resulting data sets. Uncertainty, in the form of confidence, variability, and error, as well as model bias and trends, is regularly included within data sets and is used to express descriptive, qualitative characteristics of the data. Because uncertainty is crucial in understanding the reliability of information and thus in objectives such as decision making, its absence can lead to misrepresentations and incorrect conclusions. Too often, traditional visualization approaches overlook available uncertainty information. As the importance of visualizing these large, complex data sets grows, the actual task of visualizing them becomes more complicated; incorporating the additional data parameter of uncertainty into the visualizations becomes even less straightforward. Difficulties in applying preexisting methods, additional visual clutter, and the lack of obvious visualization techniques leave uncertainty visualization an unsolved problem.

The Solution. The goal of this work is to create a summary plot that incorporates higher order descriptive statistics to concisely present data with uncertainty information. This work takes inspiration from the visual devices used in exploratory data analysis and extends their application to uncertainty visualization. The statistical measures often used to describe uncertainty are similar to measures conveyed in graphical devices such as the histogram and box plot. This research investigates the creation of the summary plot, which combines the box plot, histogram, a plot of the central moments (mean, standard deviation, etc.), and distribution fitting (Figure 27. The box plot has a canonical feel; the "signature of the plot is easily recognizable and does not need much explanation to allow for a full understanding. The focus of this work is to create a summary plot that similarly incorporates higher-order information, allowing for the quick identification of characteristic features. This higher-order signature provides at-a-glance recognition of variations from normal and allows easy comparison of data distributions in detail. In addition, a 2D extension of the summary plot has been created, which provides for the comparison of correlated (Figure 28) and ensemble data (Figure 29).

The Impact. The 1D and 2D summary plots provide a simple way to annotate features of a distribution, enhance distinguishablity between data sets, and allow for the straightforward comparison of multiple distributions. They contain, by nature, uncertainty information expressed foremost by standard deviation, but also through the higher order characteristics of the distribution. In comparison to the box plot alone, the summary plot quickly exposes salient features of the data set, such as the existence and location of outliers, the amount of variability, and the skewness of a distribution. The presentation of data in a summarized and easy to read form can quickly

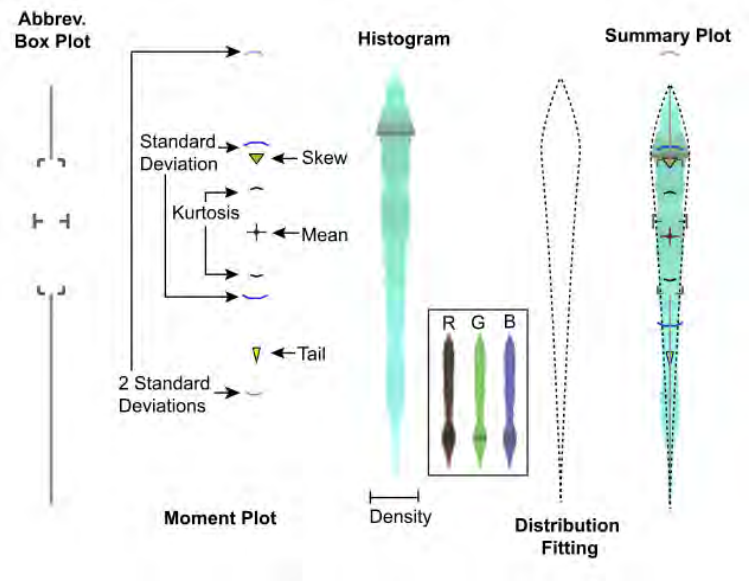


Figure 27: The summary plot which consists of an abbreviated box plot, a plot of higher order moment statistics, a density display presented using a symmetric histogram type display, and the results of distribution fitting. This plot advances the box plot to incorporate information not visible in the box plot and aids in understanding a data distribution which provides insight into the uncertainty of the data.

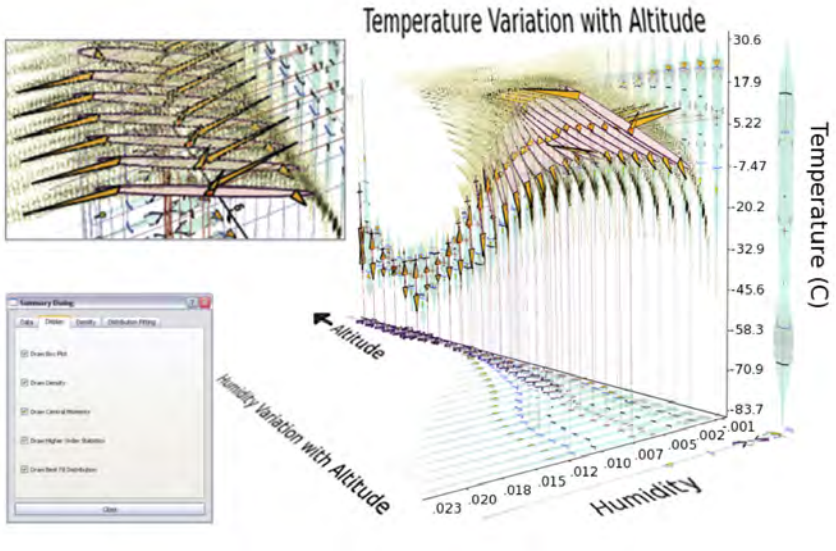


Figure 28: Right: A collection of 2D plots. Left,top: Close up of the collection. Left, bottom: User interface that helps in reducing visual clutter. Using a collection of 2D summary plots allows for the understanding of trends across correlated data sets. In addition, a user interface is provided to aid in understanding by allow the user to choose which parts of the plot to display and querying specific regions of the data.

communicate information about large amounts of data and the data's uncertainty, emphasizing meaningful characteristics and facilitating visual comparisons.

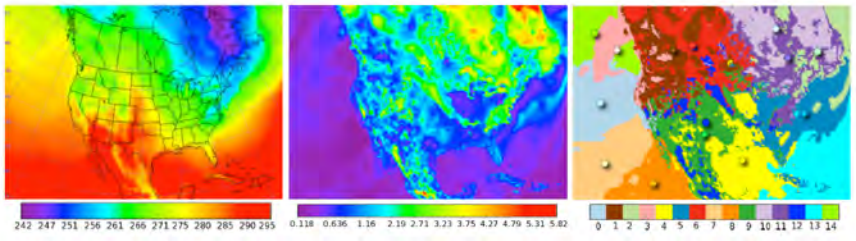


Figure 29: Mean temperature value across North America. Center: Variance of the temperature data. Right: Temperature data clustered based on variance and spatial location. Because the summary plots have a high amount of visual complexity their use across large regions of data is not feasible. However they are quite effective for understanding local areas. In this example we have an ensemble of temperature data on a fairly sparse grid across North America. We have computed the mean and variance of this data across the spatial domain. We then want to tease out interesting areas of the data, and to do so we use a clustering algorithm that groups the data both on spatial location as well as variance values. The points in the rightmost image show exemplary locations of variance values. These locations are then displayed using summary plot.

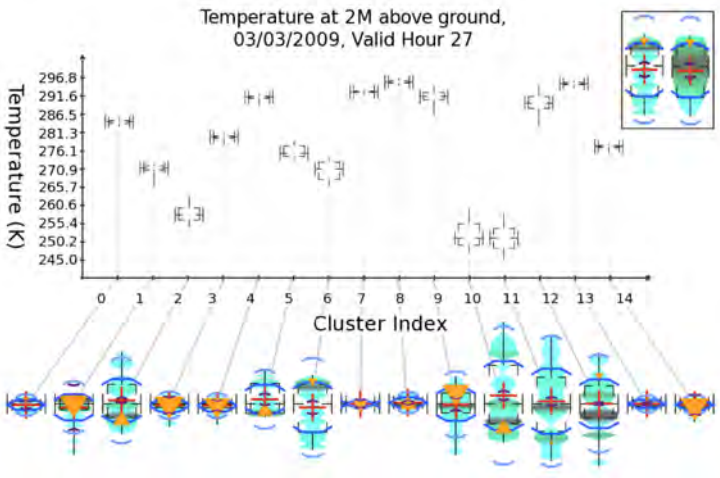


Figure 30: Summary plots to show the distribution of temperature values for cluster locations. Because the nuances of the summary plots are difficult to see in this context, this image has scaled up each summary plot and the location of each is indicated in the graph via the positions of the respective box plots. In addition, on the top right a comparison of difference density estimation techniques is shown, histogram binning on the left and kernel density estimation on the right.

3.7 Remote Collaboration Technology

Remote Collaboration

In today' highly distributed communication environment where day-to-day use of the Internet is the general norm, visualization experts are looking into the feasibility of using the web, and common web-based tools, to achieve collaboration amongst users working in different geographical locations.

From a generalized perspective, every viewer handles two kind of data sets: (1) the scientific data input to the visualization process, and (2) the meta-data that includes many application dependent parameters like the viewpoints, color-maps, transformation coordinates and so forth. The generalized difference between the two is that the input scientific data is typically static in nature whereas the meta-data is typically dynamic and that changes with every user interaction. Our collaborative system (Figure 31) targets to attain real time synchronization amongst the meta-

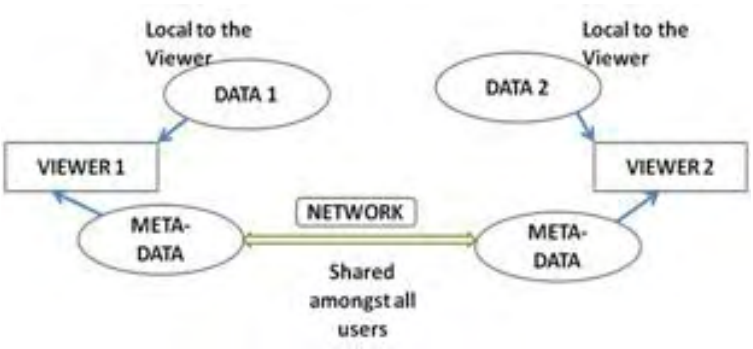


Figure 31: General architecture.

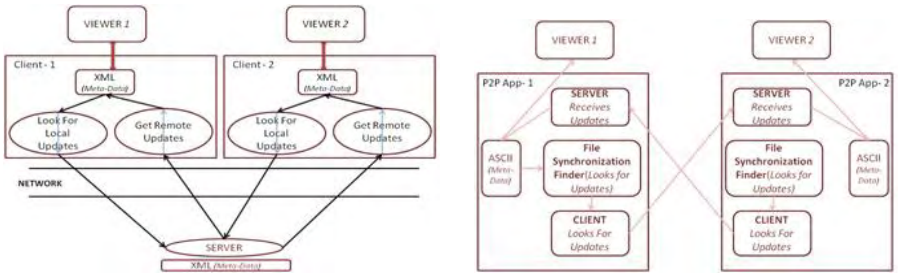
data of all the remote viewers, thus enabling remote collaboration. There is no constraint on the actual data, it can be similar at the local ends or can be different.

Potential Impact

Remote collaborative tools will largely aid people with similar research interests. Researchers using the tools can remotely interact with one other and collaborate their work. At present one of the collaborative tools, allows remote collaborations among distributed science teams commonly working in climate research.

Architecture and Models

The main issue in achieving collaboration between remote viewers is to attain a real-time synchronization of meta-data. To address this issue, we have developed two different prototype approaches: (1) web-server based model, which is more distributed and generalized and does not need to create a TCP port on the local machine; (2) a peer-to-peer model that has both a built-in client and server, and that is more suitable for LAN environments (Figure 3.7).



(a) Web server model. (b) Peer-to-peer model.

Applications

ViSUS 2D Image Viewer. The viewer allows the visualization of two dimensional images in IDX format at varying resolutions 32. Using the remote collaborative tool, one can have two or more viewers running collaboratively. Any change done to a local viewer will be made visible to the other, remotely connected viewers. This collaboration is achieved by remotely synchronizing the meta-data of the viewers. The meta-data is stored in an external file, that the viewer queries and

grabs at a regular intervals. The collaborative tool makes sure that any change in the meta-data at any local viewer is transcended to all the connected users. Two or more remote viewers can simultaneously work in tandem on their data. The system also allows any external annotations on the viewed image to be transcended across the network making it available to all the users.

Figure 32: Two remotely connected viewers.

Merge Tree Viewer. A merge tree is a topological structure that encodes a one-parameter family of features. Its structure can provide important insights into how features in large scale scientific data will change as their defining parameter changes. The Merge Tree Viewer is a tool that has been developed as part of a analysis toolkit for petascale combustion simulation data (Figure 33). The tool synchronizes with other components that generate statistics from the data and a viewer that shows the simulation over time in 3D. The toolkit provides the end user the means of studying the simulation data from multiple perspectives. Currently, the tree viewer application features the ability to provide a seamless collaborative environment for multiple users. Eventually, the aim is to add this functionality to the whole toolkit.

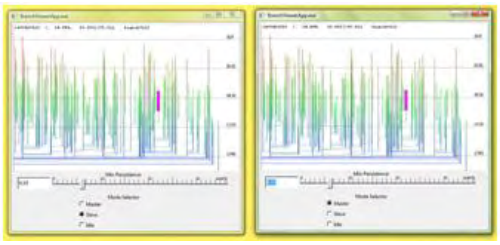


Figure 33: Two remotely connected merge tree viewers.

Remote Visualization of Climate Data. This viewer enables the visualization of climate trends of the planet earth spanning two hundred years (Figure 34). The viewer uses a smarter meta-data type: XML data structure. The advantage using the XML structure is that, it decreases the network over head, since only the pair that changes in value needs to be transmitted across the network. This project is still ongoing, so at present some of the nodes of the viewer is not written to the XML file, still one can see most of the changes on the remote viewer.

3.8 Discrete Flow Maps

The Problem. Compute topological structures on vector-valued data robustly. Show instabilities explicitly to provide more complete view of flow.

The Solution. A new data structure, discrete flow maps, used to represent flow fields on triangular meshes of 2-manifolds. Flow maps encode flow behavior across the boundary of each triangle instead of traditional vector samples at mesh vertices.

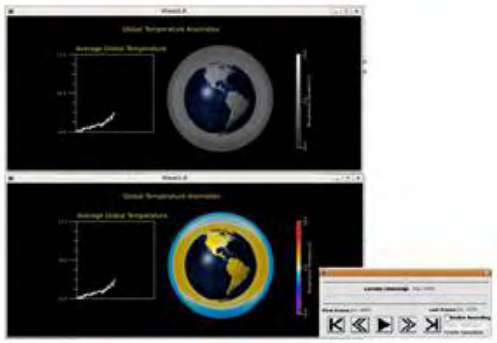
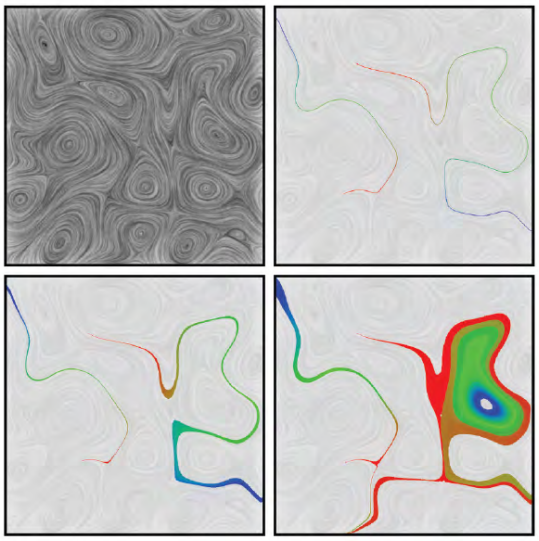


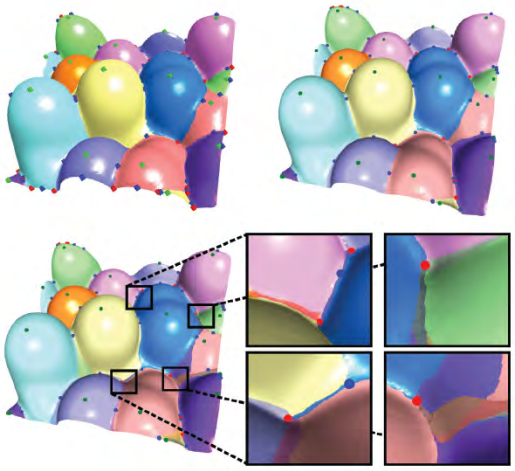
Figure 34: Two remotely connected climate data viewers.

The Impact. New views of flow fields that improve the quantification of error, complement traditional flow visualization techniques, and are a first step forward for studying topological segmentations when the data is uncertain.

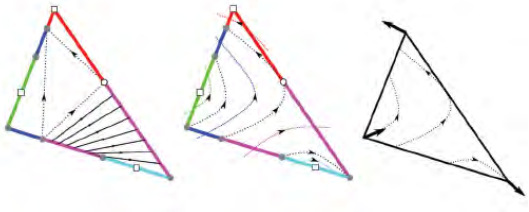
Project Summary. This work is a direct collaboration with the DOE/ASCR MAPD (Mathematics for Analysis of Petascale Data) project joint with Sandia National Labs, Texas A&M, and the University of Utah. While the basic research is being claimed under MAPD, VACET will be concerned with the deployment of new technologies produced by the effort. We anticipate the research will lead to new strategies, algorithms, and software technologies for handling petascale data by applying topological techniques that synthesize approaches for handling uncertainty.



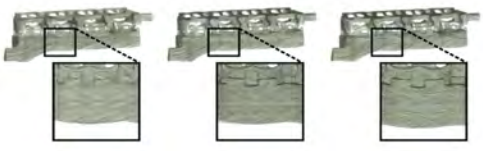
(a) Stream-fronts.



(b) Fuzzy topology.



(c) Flow maps.



(d) Vector field simplification.

Accomplishments this Period.

- Mathematical Foundations developed regarding flow maps: (1) Developed an in depth analysis of the possible flow behavior across a single triangle when piecewise linear flow is assumed. (2) Algorithm for constructing the flow map datastructure from an input mesh with sampled vectors at mesh vertices. (3) Isolated 23 topological cases for flow maps.
- Stream-fronts. (1) Using flow maps, we replace traditional notions of flow integration with a map lookup. (2) This lookup enables propagation of error terms. (3) The traditional concept of a streamline can thereby be replaced with a stream-front which propagates a fattened region forward instead of a single point.
- Fuzzy Topology. Developed topological techniques to building fuzzy views instead of the conventional crisp, clean views of the data. Using stream-fronts, we redesign topological tools to visualize stability of topological computations.
- Vector Field Simplification. Flow maps enabled a mesh compression that simplifies a vector field while preserving topological structures. We enhanced the typical edge-collapse algorithm with a measure of flow error in terms of map distortion.

4 Common Infrastructure Projects

4.1 VisIt Hero Runs

Background

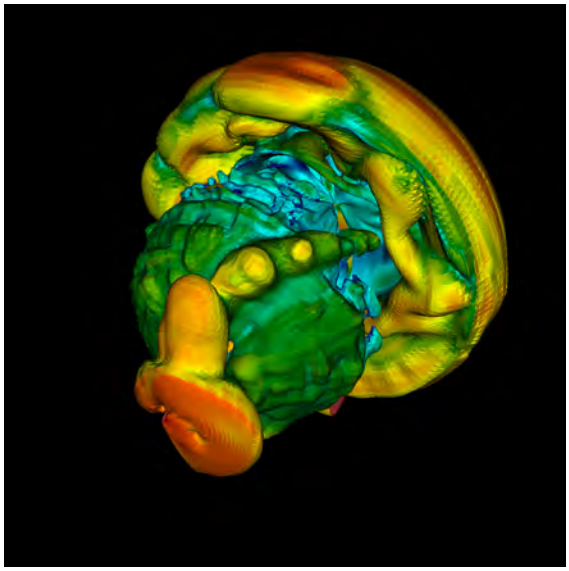
As the by-product of advances in technology is "more and more data," one issue facing the visualization and analysis community is the feasibility of using today's largest computational platforms for knowledge discovery. To gain better insight into this issue, VACET researchers recently conducted a series of experiments aimed at fostering a better understanding of functional and performance limits that might be encountered when running a production-quality visualization application at extreme levels of concurrency on data sets of unprecedented size. The results, which we discuss in this section, suggest this approach is viable and that visualization research and development efforts have produced technology that is today capable of ingesting and processing tomorrow's data sets.

Another purpose of these runs was to prepare for establishing VisIt's credentials as a "Joule code," or a code that has demonstrated scalability at a large number of cores. VisIt is the first and only visual data analysis code that is part of the ASCR Joule metric, which aims to track code performance (scalability) over a period of time.

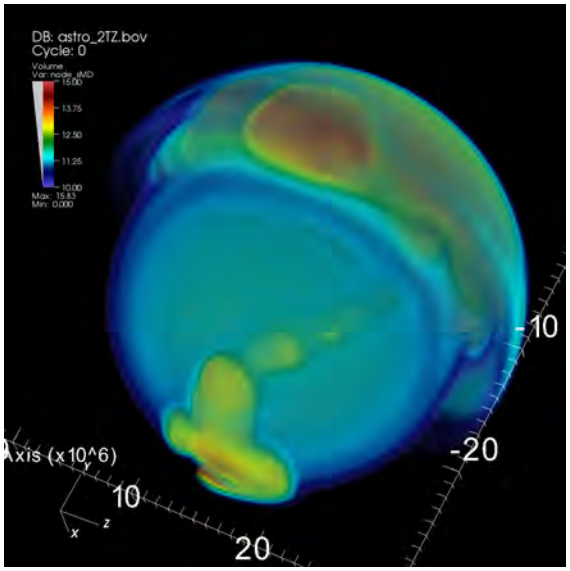
The team's experiments consisted of running the VisIt software application on several of the nation's largest computing platforms and on data set sizes ranging from 500 billion (two terabytes per scalar) to 2 trillion cells (eight terabytes per scalar) and at concurrency levels ranging from 8000 to 64,000 cores. Each experiment consisted of running VisIt in parallel: loading in data, performing two common visualization tasks (isosurfacing and volume rendering), and producing an image (Figure 35).

The Problem. One of VACET's central mission objectives is to enable visualization at the petascale. To that end, we want to ensure that production quality visualization software will be scalable to the largest data sets and on DOE's largest computational platforms. We also want to identify and understand bottlenecks.

The Solution. We embarked on a multi-platform study where we looked at very large datasets consisting of trillions of cells using ten of thousands of cores. While no such datasets of this size exist in practice today, we anticipate they will become increasingly commonplace. Therefore, our



(e) Isocontouring of two trillion zones on 32,000 Opteron cores of JaguarPF, a Cray XT5 at OLCF/ORNL.



(f) Volume rendering of two trillion zones on 32,000 Opteron cores of Franklin, a Cray XT4 at NERSC/LBNL.

Figure 35: Our functional performance experiments consist of loading extremely large data sets and executing visualization algorithms at extreme levels of concurrency producing images of isocontouring (left) and volume rendering (right).

objective is to "lead the target" by studying scalability of production-quality visualization software using datasets of a size we expect will become increasingly common.

The Impact. Our experiments show that our production-quality visual data exploration and analysis infrastructure, VisIt, is capable of processing tomorrow's datasets today on DOE's largest computational platforms. These experiments revealed bottlenecks in the application, which we have since repaired and folded into the production code that is released publicly to the worldwide scientific community.

Accomplishments this Period.

- We submitted a paper to IEEE Computer Graphics and applications describing the results of our scalability experiments. This article was accepted for publication and appears in the May/June 2010 issue.
- LLNL developer Cyrus Harrison adopted our methodology (including scripts) for LLNL's new visualization platform, `graph.llnl.gov`. His runs went as large as 12,000 cores and visualized up to eight trillion cells. This is a higher cell-to-processor ratio than we used in our previous study.

4.2 Production Quality AMR Visual Data Exploration and Analysis Infrastructure

The Problem. An increasing number of DOE simulation in various science areas (e.g., fusion, particle accelerator modeling, modeling of porous flow/carbon sequestration, astrophysics/cosmology, climate) utilize adaptive mesh refinement (AMR) simulations. An ongoing problem is the need for production-quality, parallel capable visualization software that addresses the challenges posed by data on AMR grids. Some of the challenges are so unique that some code groups have invested

their own resources into custom, home-grown visualization applications. The result is a duplication of effort along with expensive, one-off solutions that often don't have desired functionality or performance.

The Solution. We are deploying VisIt as a visualization tool that already has basic AMR support to AMR code development teams (directly) and scientists performing AMR simulations (mostly indirectly but also directly). To make possible the effective visual analysis of AMR data, we are extending VisIt's basic infrastructure for AMR data (spreadsheets, performance enhancements, bug fixes, user interface enhancements) and perform research in new AMR visualization/analysis techniques (AMR streamlines, crack-free isosurfaces, etc.) that are deployed to the science community in VisIt. This approach allows us to leverage a significant investment in VisIt to quickly deliver new AMR visualization capabilities to the science community.

The Impact. Our work will provide scientists using AMR simulations with the necessary tool to visualize and analyze their simulation results and derive insights from them. It will address an important need that arises from the proliferation of AMR-based simulation codes. In one instance, the SciDAC Applied Partial Equations Center (APDEC), our efforts have resulted in APDEC abandoning its own internal effort for creating/maintaining an AMR visualization tool and adopting VACET technology instead, thereby resulting in direct cost savings to a major SciDAC science effort in terms of software development and maintenance effort, as well as new functional and capability characteristics they need to perform their science (Figure 36).

Accomplishments this Period. During this reporting period, our team continues to engage in detail-oriented software engineering activities to implement specific features required by ADPEC along with minor bug fixes and performance optimization. We continue our interactions with APDEC as well as researchers in the LBNL Center for Computational Sciences and Engineering (CCSE) in an ongoing effort to effectively deploy VisIt to their teams and help them apply it to their specific problems. Such interactions include one-on-one consulting as well as group tutorials.

4.3 Integrating FastBit into VisIt

The Problem. In many scientific disciplines, a central challenge is the rapid visual exploration and analysis of very large, multi-dimensional, multi-variate, time-varying data. Our initial focus on this large problem space is on particle-based data, which form the basis for our work with science researchers in accelerator modeling and fusion. These science areas require efficient means for identifying relevant particle subsets (e.g., particle beams) as well as an efficient means for tracing and analyzing particles over time.

The Solution. In order to allow scientists to explore (query) extremely large, high-dimensional data we developed a novel method for rendering histogram-based parallel coordinates, which serve as main interface for defining data queries. We compute the required conditional 2D histograms efficiently using the data management system FastBit. FastBit furthermore allows us to accelerate the computation of threshold and ID-based queries. Fast computation of ID-based queries is the basis for efficient implementation of tracing of particles over time. We integrated FastBit with the visualization system VisIt providing users a production-quality software tool for visual exploration of particle data.

The Impact. Using our methods we were able reduce the time for tracing of particles over time from hours using the custom code of our stakeholder to seconds. We, therefore, enable researchers to now perform particle tracing (an essential tool for understanding the temporal behavior particles) repeatably, significantly improving the overall analysis. Using our parallel coordinates interface scientist can more accurately define particle beams, hence, improving the quality of the overall analysis. Based on this interface, scientist may also refine selections based on information from

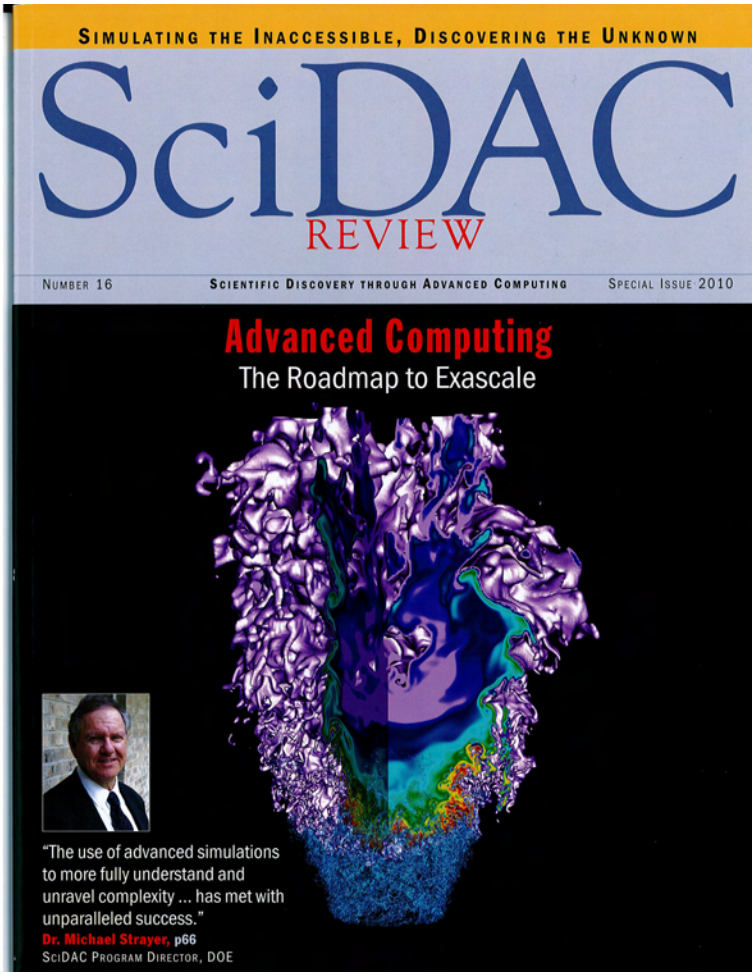


Figure 36: Simulations that ran on the Franklin Cray XT at NERSC captured the detailed structure of a lean hydrogen flame on a laboratory-scale low-swirl burner. Image courtesy of M. Day, LBNL. Visualization generated using VisIt by M. Day and A. Nonaka (LBNL CCSE) with help by G. H. Weber (LBNL Vis Group/NERSC Analytics).

multiple time points. using this new capability, researchers can perform a more detailed analysis of the data than previously possible. This work, while initially developed for and applied to problems in accelerator modeling, apply to any particle-based data, and we are presently extending this work for use in several fusion projects that use particle-in-cell based simulation codes.

Accomplishments this Period.

- Bug fixes to HDF_UC reader.
- Integrated HDF_UC capabilities into H5Part reader. The goal is a unified reader. Integration is complete. Need to do extensive testing.

4.4 Data Parallel Analysis and Graphics with R

The Problem. Statistical analyses (clustering, regression, extremes modeling, sampling, analysis of variance, etc.) are not available for large data in a form suitable for use on large-data problems and on DOE's production parallel computing infrastructure. As a result, statistical analyses are used in limited ways in large data visualization and analysis of scientific data.

The Solution. The vision is to provide production-quality, parallel capable statistical computing infrastructure on DOE's open computing platforms. Our approach is to enable the R software environment for statistical computing and graphics to be used for data-parallel analysis. We are making available data-parallel R computational methodology available to VisIt for the purpose of combining visualization and statistical analysis. We first focus on analysis of extremes and clustering, which are methods immediately relevant to out stakeholders.

The Impact. Our work enables scientists to see through the fog of variability in large data sets. Statistical methods are descriptions of variability and its underlying governing principles. We facilitate visualization at petascale and beyond by statistical selection of representative features. This is primarily accomplished through clustering of features into a smaller number of classes. A broader impact is to bring half of applied mathematics (represented by statistics) as players to the large data analysis table and develop more large data statistical analysis algorithms.

Accomplishments this Period.

Much work in this period centered around the development and scaling of a batch SPMD version of the k-means clustering code that includes cluster uncertainty. The batch version is necessary to sidestep unstable behavior of interactive Rmpi and to scale further to jaguar. The important feature of this code is that it computes cluster uncertainty based on the level of agreement between random starts. This computation accounts for expected agreement in a way that is appropriate for highly unequal cluster sizes and massive data sets. Uncertainty is the driver for determining the appropriate number of clusters for a given data set. Along the way, we also realized that sampling can be used to drastically reduce the convergence time for k-means and its implementation resulted in a factor of 10 reduction in run time. Sampling is effective for k-means because the uncertainty of a mean only depends on the size of the sample and not on the size of the data set from which the sample was taken. Current scaling results from the lens cluster are in the following table:

Cores	netCDF Read	Local File Read #1	Local File Read #2	Local File Read #3
32	18:59	12:36	11:17	11:23
64	08:17	05:44	05:47	n/a
128	05:09	03:53	n/a	n/a
256	04:59	01:50	02:08	n/a

Table 1: Scaling studies of the k-means with uncertainty Rmpi implementation on the lens cluster at ORNL. Data is in MM:SS format (minutes/seconds).

Some replicates of runs are shown to give a sense of the variability among the timed runs. Scalability is very good, particularly when only a local file is read. Parallel netCDF reading from the Lustre file system is less scalable.

The work in this period also lead to a better understanding of the infrastructure that is needed for data-parallel work in R and its connection to VisIt visualization. As a result, some of the deliverables were changed. Primarily we are now focusing on the batch runtime environment as the interactive Rmpi environment is not stable for scaling beyond 16 cores. Addressing Rmpi interactive stability seems outside the scope of VACET at this time.

We added a parallel data writer function to produce netCDF files from Parallel netCDF as the tool necessary to connect with VisIt visualization. The output of a clustering algorithm can be of the same size as the input data set, requiring parallel treatment in writing. Complex analysis algorithms such as clustering can run in batch as a preprocessing step and the results can be interactively visualized by VisIt.

We also added deliverables (wrapper functions for the BLACS communication library, wrapper functions for the PBLAS library) that will bring data-parallel matrix computation to R. This

functionality will enable easier parallelization of statistical analysis functions in R. It will connect DOE/ASCR's large investments in parallel linear algebra to R's statistical analysis capabilities.

4.5 ViSUS Core Infrastructure

The Problem. Like many long-term software projects, the source code for the main ViSUS I/O routines have become overly large and often specialized (for example, 2D and 3D data had become separate trunks in the repository). The software became increasingly more difficult to modify with new features and some needed functionality could not be easily added given previous design decisions.

The Solution. We have taken the opportunity to redesign the ViSUS I/O routines from the ground up, using "lessons-learned" from the past system to design a new infrastructure that is both lightweight and flexible for possible needs down the road. The main I/O routines now sit below an API to prevent a repeat of the bloat in the past system.

The Impact. We hope this new code and API are both easier to use and allow for larger flexibility in adding new features to the ViSUS system.

Accomplishments this Period.

- The I/O library is now written entirely in C. Some of the larger systems in the past did not support C++ therefore the ViSUS library now is more portable to VACET systems.
- There is now a documented API for the ViSUS I/O routines.
- The API now has a python wrapper to aid adding ViSUS to technologies such as Vistrails and CDAT.
- We've discarded the specialized code for 2D and 3D data. The ViSUS I/O routines are now N-dimensional.
- The ViSUS file format can now handle adaptive resolution.
- The ViSUS file format and I/O routines can now handle an arbitrary number of bits per species sample allowing for user definable precision.

Upcoming Plans.

- Discard FLTK as the user interface for the ViSUS viewer and use plain OpenGL. We hope to have a viewer that requires no additional dependencies to run other than OpenGL. By doing this we hope to have one standard viewer run on a variety of platforms including mobile devices such as the iPhone and Android.
- Create a single viewer for 2D and 3D data. Currently, these are still separate and tied to the previous ViSIS I/O routines.
- Currently, the code uses a bitmask to denote whether or not a particular location contains data. We would like to enable the I/O routines to allow for separate bitmasks per data element.
- Complete performance tests to ensure the new I/O system is equivalent to the previous version.
- Extend the ViSUS file format to handle mixed local and remote data to allow for remote viewing with local caching.

5 Publications

5.1 Peer-reviewed Journal Articles

1. H. Childs, D. Pugmire, S. Ahern, B. Whitlock, M. Howison, Prabhat, G. Weber, and E. W. Bethel. Extreme Scaling of Production Visualization Software on Diverse Architectures. *Computer Graphics And Applications, Special Issue on Ultrascale Visualization*, 30(3):22–31, May/June 2010.

2. M. Isenberg, P. Lindstrom, and H. Childs. Parallel and Streaming Generation of Ghost Data for Structured Grids. *Computer Graphics And Applications, Special Issue on Ultrascale Visualization*, 30(3):50–62, May/June 2010.

3. L. J. Gosink, C. Garth, J. C. Anderson, E. W. Bethel, and K. I. Joy. An Application of Multivariate Statistical Analysis for Query-Driven Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2010, to appear.

4. John C. Anderson, Christoph Garth, Mark A. Duchaineau, and Ken Joy. Smooth, Volume-Accurate Material Interface Reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2010, to appear.

5. Mario Hlawitschka, Christoph Garth, Xavier Tricoche, Gordon Kindlmann, Gerik Scheuermann, Ken Joy, and Bernd Hamann. Direct Visualization of Fiber Information by Coherence Direct Visualization of Fiber Information by Coherence. *International Journal of Computer Assisted Radiology and Surgery*, 5(2):125ff, April 2010.

6. Peer-Timo Bremer, Gunther H. Weber, Valerio Pascucci, Marcus S. Day, and John B. Bell. Analyzing and Tracking Burning Structures in Lean Premixed Hydrogen Flames. *IEEE Transactions on Visualization and Computer Graphics*, 16(2):248–260, March/April 2010. LBNL-2276E. doi:10.1109/TVCG.2009.69. In press.

7. Oliver Rübel, Cameron G. R. Geddes, Estelle Cormier-Michel, Kesheng Wu, Prabhat, Gunther H. Weber, Daniela M. Ushizima, Peter Messmer, Hans Hagen, Bernd Hamann, and Wes Bethel. Automatic Beam Path Analysis of Laser Wakefield Particle Acceleration Data. *IOP Computational Science & Discovery*, 2(015005 (38pp)), November 2009. LBNL-2734E.

8. M. Day, J. Bell, P.-T. Bremer, V. Pascucci, V. Beckner, and M. Lijewski. Turbulence effects on cellular burning structures in lean premixed hydrogen flames. *Combustion and Flame*, 156:1035–1045, 2009.

5.2 Peer-reviewed Conference Proceedings

1. Mark Howison, E. Wes Bethel, and Hank Childs. MPI-hybrid Parallelism for Volume Rendering on Large, Multi-core Systems. In *Eurographics Symposium on Parallel Graphics and Visualization (EGPVG)*, Norrköping, Sweden, May 2010. LBNL-3297E. **Best Paper Award winner.**

2. Mauricio Hess-Flores, Mark A. Duchaineau, Michael J. Goldman, and Kenneth I. Joy. Iterative Dense Correspondence Correction through Bundle Adjustment Feedback-based Error Detection. In *Proceedings of the International Conference on Computer Vision Theory and Applications*, page (to appear), Angers, France, June 2010.

3. Oliver Rübel, Sean Ahern, E. Wes Bethel, Mark D. Biggin, Hank Childs, Estelle Cormier-Michele, Angela DePace, Michael B. Eisen, Charless C. Fowlkes, Cameron G. R. Geddes, Hans Hagen, Bernd Hamann, Min-Yu Huang, Soile V. E. Keränen, David W. Knowles, Cris L. Luengo Hendriks, Jitendra Malik, Jeremy Meredith, Peter Messmer, Prabhat, Daniela Ushizima, Gunther H. Weber, and Kesheng Wu. Coupling Visualization and Data Analysis for Knowledge Discovery from Multi-dimensional Scientific Data. In *Procedia Computer Science*, page (to appear). Elsevier, June 2010.

4. K. Potter, J.M. Kniss, R. Riesenfeld, and C.R. Johnson. Visualizing summary statistics and uncertainty. In *Proceedings of Eurographics/IEEE-VGTC Symposium on Visualization 2010*, page (to appear), 2010.

5. Jeremy S. Meredith and Hank Childs. Visualization and analysis-oriented reconstruction of material interfaces. In *Proceedings of Eurographics/IEEE Symposium on Visualization (EuroVis)*, Bordeaux, FR, June 2010. To appear.

6. Gunther H. Weber, Sean Ahern, E. Wes Bethel, Sergey Borovikov, Hank R. Childs, Eduard Deines, Christoph Garth, Hans Hagen, Bernd Hamann, Kenneth I. Joy, Daniel Martin, Jeremy Meredith, Prabhat, Dave Pugmire, Oliver Rübel, Brian Van Straalen, and Kesheng Wu. Recent advances in visit: Amr streamlines and query-driven visualization. In *Numerical Modeling of Space Plasma Flows: Astronum-2009 (Astronomical Society of the Pacific Conference Series)*, 2010. LBNL-3185E. To appear.

7. Kristin Potter, Andrew Wilson, Peer-Timo Bremer, Dean Williams, Charles Doutriaux, Valerio Pascucci, and Chris R. Johhson. Ensemble-Vis: A Framework for the Statistical Visualization of Ensemble Data. In *IEEE Workshop on Knowledge Discovery from Climate Data: Prediction, Extremes, and Impacts, to appear.*, pages 233–240, 2010.

8. Dave Pugmire, Hank Childs, Christoph Garth, Sean Ahern, and Gunther H. Weber. Scalable Computation of Streamlines on Very Large Datasets. In *Proc. Supercomputing SC09*, Portland, OR, USA, November 2009. LBNL publication number pending.

9. Peer-Timo Bremer, Gunther H. Weber, Julien Tierny, Valerio Pascucci, Marcus S. Day, and John B. Bell. A Topological Framework for the Interactive Exploration of Large Scale Turbulent Combustion. In *Proceedings of the 5th IEEE International Conference on e-Science*, pages 247–254, Oxford, UK, December 2009. LBNL-3183E.

10. Eduard Deines, Gunther H. Weber, Christoph Garth, Brian Van Straalen, Sergey Borovikov, Daniel F. Martin, and Kenneth I. Joy. On the Computation of Integral Curves in Adaptive Mesh Renement Vector Fields. In *Dagstuhl Seminar Scientific Visualization 09 (09251)*, Wadern, Germany, 2009. In review.

11. Emanuele Santos, Julien Tierny, Ayla Khan, Brad Grimm, Lauro Lins, Juliana Freire, , Valerio Pascucci, Claudio Silva, Scott A. Klasky, Roselyne D. Barreto, and Norbert Podhorszki. Enabling advanced visualization tools in a simulation monitoring system. In *Proceedings of the 5th IEEE International Conference on e-Science*, pages 358–365, December 2009.

5.3 Invited Articles

1. Christoph Garth, Eduard Deines, Kenneth I. Joy, E. Wes Bethel, Hank Childs, Gunther Weber, Sean Ahern, Dave Pugmire, Allen Sanderson, and Chris Johnson. Twists and Turns:

Vector Field Visual Data Analysis for Petascale Computational Science. *SciDAC Review*, 15:10–21, Winter 2009. LBNL-2983E.

5.4 Book Chapters

1. E. Wes Bethel, Hank Childs, Ajith Mascarenhas, Valerio Pascucci, and Prabhat. Scientific Data Managment Challenges in High Performance Visual Data Analysis. In Arie Shoshani and Doron Rotem, editors, *Scientific Data Management: Challenges, Existing Technology, and Deployment*. Chapman & Hall/CRC Press, December 2009. LBNL-1449E.

5.5 Theses and Dissertations

1. E. Wes Bethel. *High Performance Visualization*. University of California, Davis, March 2010.
2. Oliver Rübel. *Linking Automated Data Analysis and Visualization with Applications in Developmental Biology and High-energy Physics*, volume 28 of *Schriftenreihe Informatik*. Der Dekan (hrsg), Fachbereich Informatik, Technische Universität Kaiserslautern, December 2009.

5.6 Technical Reports

1. E. Wes Bethel. Using wesBench to Study the Rendering Performance of Graphics Processing Units. Technical Report LBNL-3025E, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, 94720, 2010.

6 Outreach and Service

6.1 Outreach

Symposia and Workshops

- SIAM Conference on Parallel Processing for Scientific Computing (PP10) mini-symposium: "The Challenges Ahead for Visualizing and Analyzing Massive Data Sets." 26 February 2010, Seattle WA. H. Childs (organizer), D. Pugmire, P.-T. Bremer (and others).
- Scalar Topology in Visual Data Analysis tutorial at the IEEE Visualization 2009 Conference, Gunther H. Weber, Peer-Timo Breme, Attila Gyulassy, Hamish Carr. Atlantic City, NJ, 2009.
- International Research Training Group. K. Joy, C. Garth, E. Deines, C. Hansen, B. Hamann.

Invited Talks

- Hank Childs, Petascale Visualization Concerns on I/O & VisIt Overview, NSF Workshop on Petascale I/O, Austin, TX, March, 2010.
- Hank Childs, Extreme Scaling of Production Visualization Software on Diverse Architectures, Park City, UT, April, 2010.
- Sean Ahern, Panel: Challenges in Large Data Visualization: A Visualization Community Call to Action, Atlantic City, NJ, October 2009.
- E. Wes Bethel. MPI-hybrid Parallelism for Volume Rendering on Large, Multi-core Systems. DOE Computer Graphics Forum, Park City UT, April 2010.
- E. Wes Bethel. 3D Bilateral Filtering on the GPU. DOE Computer Graphics Forum, Park City UT, April 2010.

6.2 Service

Conference Chair

- DOE 2010 SciDAC Conference – Visualization co-Chairs.

Program Committee

- IEEE Visualization 2010.
- ACM/IEEE Supercomputing 2010.
- SIBGRAPI 2010 Conference on Graphics, Patterns and Images.
- 6th International Symposium on Visual Computing ISVC10.
- 2010 Workshop on Emerging Computational Methods for the Life Sciences (at HPDC 2010).
- IASTED International Conference on Computer Graphics and Imaging (CGIM 2010).
- EuroVis (Eurographics/IEEE Symposium on Visualization) 2010.
- 2010 Conference on Interactive 3D Graphics and Games.

Reviewer

- DOE SBIR Program, 2010.
- DOE/ASCR unsolicited proposals.
- EuroVis (Eurographics/IEEE Symposium on Visualization) 2010 Technical Program.
- IEEE Visualization/InfoVis/VAST 2010 Technical Program.
- IEEE Transactions on Visualization and Computer Graphics.
- International Journal of High Performance Computing Applications.

Advisory Panels

- NSF Blue Waters Visualization Advisory Committee.
- Member External Advisory Board, NIH Center for Biomedical Computing.

6.3 Awards

1. **Best Paper Award.** Mark Howison, E. Wes Bethel, and Hank Childs. MPI-hybrid Parallelism for Volume Rendering on Large, Multi-core Systems. In *Eurographics Symposium on Parallel Graphics and Visualization (EGPVG)*, Norrköping, Sweden, May 2010. LBNL-3297E.